

REVEALING AND RESOLVING CONTRADICTORY
WAYS TO REDUCE SELECTION BIAS TO ENHANCE
THE VALIDITY OF CAUSAL INFERENCES FROM
NON-RANDOMIZED LONGITUDINAL DATA

By

HUA LIN

Bachelor of Science in Applied Physics
Jinan University
Guangzhou, Guangdong
China
2001

Master of Science in Human Development and Family
Science
Oklahoma State University
Stillwater, OK
2015

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
July, 2018

REVEALING AND RESOLVING CONTRADICTORY
WAYS TO REDUCE SELECTION BIAS TO ENHANCE
THE VALIDITY OF CAUSAL INFERENCES FROM
NON_RANDOMIZED LONGITUDINAL DATA

Dissertation Approved:

Robert Larzelere, Ph.D

Dissertation Adviser

Laura Hubbs-Tait, Ph.D

Isaac J. Washburn, Ph.D

Amanda Baraldi, Ph.D

ACKNOWLEDGMENTS

I have enjoyed my graduate college experience, an intense time of learning which added more than knowledge for me. I would like to reflect on the people who have supported and helped me so much throughout this period.

My deepest thanks and admiration belong to Dr. Robert Larzelere, my advisor and mentor. You have patiently, humbly, and tirelessly supported me and walked me through each step of my graduate studies. I especially thank Dr. Laura Hubbs-Tait, my extra mentor, for her unwavering support, mother-like care, and guidance during my graduate studies. I thank my committee members, Dr. Amanda Baraldi and Dr. Isaac Washburn, for their commitment to my academic progress and their invaluable input into this research. I could not have asked for a better team.

I thank my husband Wakun for all the duties you provide for our family and your encouragement in this journey with me.

I thank my children who inspire my research in parenting and child development and keep me laughing.

I thank my parents, siblings, and friends for their steadfast support and love.

Finally, I thank myself for never giving up my dreams and sedulously working hard for self-improvement.

Name: Hua Lin

Date of Degree: AUGUST, 2018

Title of Study: REVEALING AND RESOLVING CONTRADICTIONARY WAYS TO
REDUCE SELECTION BIAS TO ENHANCE THE VALIDITY OF
CAUSAL INFERENCES FROM NON-RANDOMIZED
LONGITUDINAL DATA

Major Field: HUMAN DEVELOPMENT AND FAMILY SCIENCE

Abstract: The purpose of the current study is to understand the mechanism of selection bias in nonrandomized studies by exploring possible reasons for inconsistent and biased results between the simple gain scores and the residual gain scores from the human development perspective. Specifically, I test several possible reasons that lead to Lord's paradox (contradictory results from the two approaches), and test alternative models for solving the paradox to get consistent and hopefully less biased results using simulated data and data on treatments for mothers' depression from the Fragile Families (FFCW) longitudinal dataset. Monte Carlo simulations, based on Lord's paradox and reversed Lord's paradox, generated 97 conditions of 1000 simulated datasets by varying violations of assumptions of ANCOVA and simple gain score analyses. The FFCW data include two types of treatments at Wave 4 and mothers' depression severity at Wave 2, 3, 4, and 5. Seventeen covariates measured before treatment were used for propensity score matching. An alternative model, group-centered ANCOVA, was developed to compare results to another alternative model, a combination of mixture model and propensity score matching. The results indicated that Lord's paradox exists when the mean pretest score differs for the treatment group and the comparison group, and consistent results can be reached when that mean pretest difference is removed. Moreover, group-centered ANCOVA approach and the combination of mixture modeling and matching on either the pretest or the propensity score could be used to remove the pretest difference between the treatment group and comparison group. These two methods of removing the mean pretest difference between the two treatment conditions result in two distinct sets of consistent results, which are nearly as inconsistent with each other as the original inconsistency. Consistent results do not guarantee an unbiased result. This study helps to understand alternative ways to adjust for selection bias in nonrandomized studies. Under some conditions models based on simple gain scores may be less biased than models based on residual gain scores. When simple gain score are less biased, group-centered ANCOVA provides more statistical power than traditional analyses and other covariates can be added to minimize other suspected confounds.

TABLE OF CONTENTS

LIST OF TABLES	VIII
----------------------	------

LIST OF FIGURES	X
-----------------------	---

Chapter	Page
INTRODUCTION	1
REVIEW OF LITERATURE	7
Making Casual Inferences in Human Development Studies: Methods and Problems	7
The Problem of Selection Bias	7
Two “Adjusting” Approaches – Simple Gain Score and Residual Gain Score	9
Lord’s Paradox and Empirically Inconsistent Results	15
Understanding Implications of Lord’s Paradox	17
Questions Raised from Lord’s Paradox	17
Making Valid Casual Inferences	20
Assumptions	23
Another Issue: Under-Adjustment	33
Possible Solutions	35
The Latent Class Growth Model	36
The Group-Centered ANCOVA	37
The Propensity Score Adjustment Approach	39
Current Study	41

STUDY I: SIMULATION.....	49
Methods.....	50
Testing Research Question 1: Simulating Lord’s Paradox and Reversed Lord’s Paradox	50
Testing Research Question 3: Simulating Violations of ANCOVA Assumptions.....	52
Testing Research Question 4	55
Results.....	57
Research Question 1	58
Research Question 3	58
Research Question 4	62
Discussion	63
STUDY II: THE FRAGILE FAMILY AND CHILD WELL-BEING DATA.....	70
Methods.....	72
Participants.....	72
Measures	72
Results.....	76
Research Question 1	78
Research Question 2	79
Research Question 3	80
Research Question 4	88
Discussion.....	94
GENERAL DISCUSSION	97
Summary and Interpretation of Results	97
Model Comparisons	105
ANCOVA vs. Propensity Score (Combining or not Combining Mixture Modeling)	106
Simple gains Score vs. Residual Gain Score	107
Simple Gain Score vs. Group-Centered ANCOVA.....	112

Directions for Human Development Studies	114
Strengths	115
Limitations and Future Suggestions.....	116
REFERENCES	118
APPENDIX A.....	169
APPENDIX B	174
APPENDIX C	183
APPENDIX D.....	185

LIST OF TABLES

Table	Page
1. Research Question Summary.....	130
2. Simulation Results on Paradox Varying Distribution, Slope, and Standard Deviation.....	131
3. Simulation Results on Varying Pretest and Posttest Different for Females	133
4. Simulation Results on Lord’s Paradox and Reversed Paradox for Comparing Difference Approaches	135
5 Simulated Data for Comparing Results from Original and Matched Data.....	136
6 Frequency of Mothers’ depression severity score	137
7 Adjusted Effects and Correlation between Covariates, Treatments, and Outcomes	138
8 Effect Sizes for Psychological and Medication Treatment for Mother’s Depression.....	140
9. Linear Growth Model Using a Zero-Inflated Hurdle Model to Estimate Trajectory Groups.....	141
10. Three Trajectory Groups Estimates Based on Zero-Inflated Hurdle model	142
11. Comparing Results between the Simple Gain Score Approach and the Residual Gain Score Approach from Original and Matched Data within Three Sub-groups for FFCW Data	143
12. Mean and Propensity of Being in the Treatment Group for the 16 covariates (Depression at Wave 4).....	145

Table	Page
13. Covariates Balance Based on Standardized Differences (Cohen's d), Before and After Matching within Three Trajectory Groups Using Samples Including Zeros Depression at Wave 4 and 5.....	146
14. Covariates Balance Based on Standardized Differences (Cohen's d): Before and After Matching within Three Subgroups Using Samples NOT Including Zeros Depression at Wave 4 and 5.....	147

LIST OF FIGURES

Figure	Page
1. Van Breukelen's (2013) Figure on ANCOVA Adjustment of the Posttest Difference for the Pretest Difference	148
2. Lord's (1967) Paradox	149
3. Propensity score distribution for simulated Lord's paradox data (OLPD) and the reversed Lord's paradox data (RLPD).....	150
4. Matched and unmatched samples for simulated Lord's paradox data (OLPD) and the reversed Lord's paradox data (RLPD).....	151
5. Histogram plots for residuals in normal distributed and non-normal distributed data	152
6. CIDI-SF Questions and Frequency on Wave 2.....	153
7. Histogram plots on the frequency of mothers' depression severity.....	154
8. Density Plots for Imputed and Original Data	155
9. Cross-Lagged Panel Model of Mothers' Depression Featuring Wave 3, 4, and 5	156
10. Latent Growth Model of Mothers' Depression Featuring Wave 3, 4, and 5.	157
11. Histogram plots for the distribution of residual of depression at Wave 5 regression on depression on Wave 4 and the two type of treatment when testing research question 3.3.....	158
12. Distribution for mothers' depression at Wave 5 for each unmatched trajectory subgroup.....	159
13. Histogram plot for the distribution of residual of depression at Wave 5 within each trajectory subgroup: regression on Wave 4 depression and the two type of treatment when testing Research Question 3.6	160

14. Compare propensity scores between the treatment and control group using unmatched sample: propensity score calculated based on pretest scores only.....	161
15. Compare propensity scores between the treatment and control group using unmatched sample: propensity score calculated based on multiple covariates.....	162
16. Compare propensity scores between the treatment and control group using matched sample: propensity score calculated based on pretest score only	163
17. Compare propensity scores between the treatment and control group using matched sample: propensity score calculated based on multiple covariates.....	164
18. Distribution for mothers' depression at Wave 5 for matched samples within each trajectory group	165
19. Histogram plots for the distribution of residual of depression at Wave 5 within each trajectory subgroup using matched samples based on pretest outcome as the only covariate: regression on Wave 4 depression and the two type of treatment.....	166
20. Histogram plots for the distribution of residual of depression at wave 5 within each trajectory subgroup using matched samples based on multiple covariate: regression on Wave 4 depression and the two type of treatment	167

CHAPTER I

INTRODUCTION

Human development research is fundamentally the study of change (explaining between-person differences in within-person changes). Its first goal is to describe changes. The second goal is to explain changes. The third and ultimate goal is to promote optimal human development, which requires making valid causal inferences, interpreting the relationship between the cause and the effect, in response to the second goal.

Valid explanations require more than accurate predictions. Epidemiologists make a crucial distinction between non-causal risk factors, such as hospitalizations, and causal risk factors, such as smoking (Kraemer et al., 2001). Overnight hospitalizations during the year are correlated with poorer physical health at the end of that year ($r = -.33$; pp. 12-15; Angrist, 2009), whereas smoking makes people 15 to 30 times more likely to get lung cancer (Centers for Disease Control and Prevention, 2017). Unlike smoking, however, hospitalization is a non-causal risk factor. Avoiding smoking will reduce one's probability of getting lung cancer, but avoiding hospitalizations will not improve one's health, despite similar adverse longitudinal associations. Life-threatening illness is the

major cause of the poor health outcomes after hospitalization. Medical treatments requiring hospitalization are designed to improve that poor prognosis, but they do not always eliminate the poor prognosis completely. Thus, hospitalization predicts, but does not cause poor health afterwards, whereas smoking can cause lung cancer. This analogy could be extended to understand studies in human development. Although human development studies do not clearly distinguish non-causal risk factors and causal risk factors, non-causal factors and causal factors are common but mixed together in predictions used to make casual inference. For instance, frequent destructive marital conflict and participating in marital therapy both predict divorce, but frequent destructive marital conflict increases the probability of divorce, whereas participating in marital therapy is a corrective action for helping to reduce marital problems. If we use the results of prediction to interpret causation naively, both frequent destructive marital conflict and participating in marital therapy can be seen as increasing the probability of divorce, which means that people should not go to a marital therapy intervention program when marriage relations are in a tough situation. Apparently, that is incorrect. These examples show that simple prediction should not be used to make causal inference. Making valid casual inferences depends on more rigorous analyses than simple predictions. The analyses need to be rigorous in terms of the requirements for making casual inference that a cause A and the effect B are not only associated with each other, but also that the cause A must be the ONLY interpretation for the effect on B, and if the cause A did not happen the effect B would not happen. Taking the divorce example, couples who are going to marriage therapy should have more probability to get divorced than couples who do not need marriage therapy. This does not mean that going to marriage theory could cause

divorce. If marriage therapy could cause divorce, then those who have gone to marriage therapy and finally got divorced would not have divorced if they had not gone to marriage therapy. Apparently, this is not the case that those couples that went to therapy and finally divorced would not get divorced if they had not gone to therapy. Thus, theoretically, the relationship between marriage therapy and divorce do not meet the requirements for making causal inference. Presumably, it is the marital problems that led the couple to go to therapy that cause their greater likelihood of divorce, and we assume that marital therapy reduces that likelihood of divorce, but not to zero.

The question is how to test this theory that meets the requirement of making causal inference. The consensus is that randomized studies, e.g., randomly assigning participants to a treatment group and a comparison group, represents the gold standard for making valid causal inferences. They are required before new prescription medications can be approved in the United States. They are required for the vast majority of meta-analyses in health care (Reeves, Deeks, Higgins, & Wells, 2008), which also may account for the fact that scientific advances are much more evident in medical practice than in human development. Where possible, randomized studies ensure that the comparison group is equivalent to the treatment group on all possible confounding variables, so that the only difference in the outcomes of the two groups is due to the treatment effect. From the human development perspective, researchers are interested in how within-person change due to treatment could lead to between-person differences. Most of the current research methods analyze the effect of treatment on between-person differences to get the treatment effect. The problem is that the contribution of between-person differences could be due to numerous known and unknown factors. It is hard to tell which part of

between-person differences is due to treatment. Through random assignment before the treatment is introduced, the between-person difference (between the comparison group and the treatment group) is minimized; after the treatment is introduced, changes that result in between-person differences (between the treatment group and the comparison group) would be expected to be due to the treatment. Note that, in this study, the difference between the treatment group and the comparison group refers to between-person differences.

However, most human developmental studies involve participant self-selection that could not be manipulated in a randomized experiment for ethical or practical reasons, resulting in *selection bias*, the differential prognoses of the comparison groups (i.e., the between-person difference) being compared for reasons such as pretest scores on the outcome, social status, education level, family history, and health conditions, other than the treatment differences. Selection bias produces biased results, which can lead to misleading interpretations of the outcomes. In order to minimize selection bias, human development investigators typically “adjust” for pre-treatment differences on the outcome variable in one of two ways: the simple gain score approach (e.g., differences-in-differences or CHANGE as in repeated measures ANOVA) and the residual gain score approach (e.g., ANCOVA or multiple linear regression). The purpose of “adjusting” for pre-treatment differences is to minimize those between-person differences that could affect treatment assignment and the outcome before the treatment is introduced. Results from these two approaches are consistent in randomized studies (Van Breukele, 2013). However, Lord’s (1967) paradox showed that these two adjustment methods could contradict each other in nonrandomized studies. Although Lord’s paradox has been

discussed for over 50 years as to which approach is unbiased, it is not clear that either of his analyses is unbiased, nor is it clear what contributes to biased causal estimates. The lack of guidance for which method is less biased results in researchers choosing a favorite statistical method to meet their research needs. Since these two approaches are building blocks for more complicated models such as cross-lagged panel analyses and latent growth models, the problem of contradictory and biased estimates based on the two baseline-adjustment approaches could extend to biased estimations when using advanced statistical models. However, little is known about whether or how Lord's paradox applies to more complicated statistics models (cross-lagged panel model and latent growth model) and what mechanisms could effectively reduce selection bias, resulting in better approximations of valid causal estimates.

The current study simulated Lord's (1967) paradox and reversed Lord's paradox, investigated problems that lead to Lord's paradox, and sought to promote valid causal inferences by comparing several models to try to resolve it. Lord's paradox reflects the contradictory results that occur when the data fit the null hypothesis of the simple gain score approach. This indicates it is also important to examine what the contradictory results would be when the data fit the null hypothesis of ANCOVA, which I called reversed Lord's paradox, because the ANCOVA approach is used in many, if not most longitudinal analyses in developmental research, e.g., cross-lagged panel analyses, hierarchical regression. In addition, although general conclusions about which approach (the simple gain score approach and the residual gain score approach) is unbiased may be impossible in nonrandomized studies, consistent results from the two approaches, which could be reached in randomized studies, may strengthen the evidence for valid causal

inferences. Thus, whether the results are consistent or not from the two approaches could be revealed by identifying problems that lead to Lord's (1967) paradox and by examining the effectiveness of the introduced models. Note that, in nonrandomized studies, consistent results may exist but could both be biased, which requires multiple model comparisons to examine whether the consistent results from one approach are unbiased or not. Thus, the present study will use these two approaches in Study I, using simulation data, to reveal and detect the problems that lead to Lord's (1967) paradox, and in Study II, using data from the Fragile Families and Child Wellbeing (FFCW) data on treatments for depression in mothers, to apply the results from the simulations and to conduct model comparisons for reaching consistent and less biased results. There is evidence from randomized studies or meta-analyses that medication treatment significantly reduces depression and that psychological treatment may or may not significantly reduce depression, but neither is harmful in increasing depression symptoms (Andersson & Cuijpers, 2009; Cuijpers, Van Straten, & Smit, 2006; Müller et. al., 2006; Parekh, 2017). If our analyses on the two types of treatment for depression in FFCW data show results that contradict the randomized studies, the results may be biased and thus deviate from the true average effect, when analyzing that particular type of change score.

CHAPTER II

REVIEW OF LITERATURE

Making Casual Inferences in Human Development

Studies: Methods and Problems

The Problem of Selection Bias

Human development investigators are interested in research on within-person changes and between-person differences in those changes and on what causes those differential changes. Valid causal inferences are necessary to use the research to promote optimal human development. To explain what causes the change involves making casual inferences, the relation between a cause and its consequence, such as whether inter-parental conflicts will increase child aggressive behaviors, whether speaking two languages at home has benefits for child cognitive development, and whether co-parenting helps children from post-divorce families adjust for parental separations. As mentioned in the previous chapter, making valid causal inferences involves more than predictions. Participating in marital therapy may predict divorce because unhappy couples may be more likely to be in the intervention program, but the intervention

program is not the cause of divorce. Marital conflict also predicts divorce and marital conflict is more likely to be a reason for divorce. Thus, a simple prediction analysis cannot be used for making valid causal inference. The reason is prediction only tells how the two variables are associated with each other, but valid causal inferences rely on unbiased estimation from correct research methods, which requires steps beyond correlations and predictions. A valid causal inference requires no pretest differences between the treatment group and the comparison group before the treatment is introduced, so that after the treatment is introduced, the group difference on the outcome is only due to the treatment. From the human development perspective, a valid causal inference requires no between-person difference before the treatment is introduced; when this condition is met prior to treatment, the between-person differences after the treatment is introduced will be only due to the treatment effect on within-person change. Thus, the treatment effect could be estimated through analyzing the between-person difference or the group difference on the outcome. The consensus is that randomized studies represent the gold approach for making valid causal inferences because it is assumed that through experimental manipulation the experimental groups are matched in all known and unknown conditions, such as their pretest scores on many variables, including the outcome score prior to getting any treatment, and the only difference between the two groups is the treatment effect.

However, most human developmental studies involve participant self-selection that could not be manipulated in experiments due to ethical reasons or practical difficulties. When studying the effect of domestic violence on individuals' depression, since domestic violence is not experimentally induced, we cannot randomly assign one

group of couples to perform domestic violence and the other group of couples to avoid domestic violence. Children living in low social economic status families may be more likely to be self-selected in the treatment group for a delinquency prevention program than children living in middle or upper class families. When the effect of the Head Start program on children's school performance is studied, typically, the treatment group includes children living in low SES families whereas the comparison group are those living in middle or upper class families. Similarly, since divorce involves couples' personal choice we cannot randomly assign couples to be in the divorced group or non-divorced group. Most of the time, people are naturally in the situation or choose to be in the situation, depending on their family environment, previous experience, and personal beliefs. Natural occurrences or participant self-selections are associated with preexisting conditions, such as families' SES, mothers' age at marriage, and parents' education, which are not matched between the active treatment group and the comparison group. Unmatched preconditions could also be used to interpret group differences as well as using the treatment to interpret the group differences. This leads to selection bias (Shadish, Cook, & Compbell, 2002; Tripepi, Jager, Dekker, & Zoccali, 2010) producing differential outcomes due to reasons in addition to the treatment, such as unmatched preconditions, which complicates making valid causal inferences (Heckman, 1979; Heckman 1990; Heckman 2010; Heckman, Ichimura, Smith, & Todd, 1998).

Two “Adjusting” Approaches – Simple Gain Score and Residual Gain Score

So, how should analyses balance out the effect of preexisting differences between the comparison group and the active treatment group? Researchers believe that preexisting conditions could include all kinds of variables that lead to a difference

between the active treatment group and the comparison group. Many human development investigators believe that “controlling for” pre-treatment differences can minimize selection bias (Diaz & Handa, 2006), and many analyses are based on the assumption that the most important pre-treatment difference to “control for” is the pre-test score on the outcome variable. “Controlling” for the pre-test may “control for” other pre-treatment differences if the effects of those pre-treatment differences affect the pre-test score as much as the post-test score. More precisely, “controlling for” the pre-test score will “control” adequately for any other pre-treatment difference that does not predict the outcome beyond the extent to which it predicts the pre-test score on that outcome. In other words, after “controlling for” the pretest score, the only pre-treatment conditions that can bias the results are those that influence change in the outcome variable over and above the extent to which it influences pre-test scores on that outcome. Most longitudinal analyses incorporate one of two basic strategies to adjust for these initial differences: the simple gain score approach and the residual gain score approach (Huitema, 2011; van Breukelen, 2013), based loosely on Rubin’s Causal Model (1974, 2004, 2005, 2006).

The simple gain score approach could be considered a more straightforward application of Rubin’s causal model (1974, 2004, 2005, 2006), which defines a treatment effect as the difference between an outcome after an active treatment and what that person’s outcome would have been if they had been in the comparison condition instead. In randomized studies, the treatment effect is the difference between the average active treatment outcome $\bar{Y}(E)$ and the average control outcome $\bar{Y}(C)$, assuming no pre-condition differences due to random assignment. When pretest scores are different, it may be more reasonable to adjust for those pretest scores in some way rather than estimate the

treatment effect by the differences between the average outcome scores by themselves. If the correct adjustment occurs with the simple gain score, we have the adjusted average treatment effect:

$$\bar{Y}_2(E)_{ads} - \bar{Y}_1(C)_{ads} = (\bar{Y}_{21} - \bar{Y}_{20}) - (\bar{Y}_{11} - \bar{Y}_{10})$$

where, $\bar{Y}_2(E)_{ads}$ is the average gain score for the active treatment group; \bar{Y}_{21} is the average outcome score measured after being treated in the experimental group; \bar{Y}_{20} is the average outcome score measured before being treated in the experimental group; $\bar{Y}_1(C)_{ads}$ is the adjusted average gain score for the comparison group; \bar{Y}_{11} is the average outcome score measured after being in the comparison group; and \bar{Y}_{10} is the average outcome score measured before participating in the comparison group. The formula could be revised to be:

$$\bar{Y}_2(E)_{ads} - \bar{Y}_1(C)_{ads} = (\bar{Y}_{21} - \bar{Y}_{11}) - (\bar{Y}_{20} - \bar{Y}_{10}) \quad (1)$$

These two formulas indicate two possibilities of no treatment effect: when the mean gain score for active treatment $\bar{Y}_1(E)_{ads}$ and the mean gain score for the comparison condition $\bar{Y}_1(C)_{ads}$ are the same, or when the pretest difference between the active treatment group and the comparison group and the posttest difference between the active treatment group and the comparison group are the same.

The other approach is the residual gain score approach, also called Analysis of Covariance for a categorical treatment (ANCOVA; Porter & Raudenbush, 1987; Reichardt, 1979) or multiple regression for a continuous treatment. Many human development investigators use ANCOVA to include the pre-test score on the outcome variable as the main covariate to “control for” in the regression formula. The idea of

regression/ANCOVA is to account for the variance of the outcome variable in terms of the unique associations of its predictors. The standardized regression coefficient is related to the percent of the variance of the outcome that could be explained by that predictor. When covariates are included in the regression formula, part of the variance of the outcome variable is explained by the covariates. Confounding effects, third variables associated with the treatment that influence the outcome, are partialled out to the extent that they are measured adequately by the covariates. Adapting the idea of ANCOVA, Campbell includes the pretest outcome variable in the regression formula as a predictor for controlling for the pretest difference between the comparison group and the active treatment group (1975). The formula is as follows:

$$Y_{ij1} = \beta_0 + \beta_1 X_j + \beta_2 Y_{ij0} + e_{ij} \quad (2)$$

Here, Y_{ij1} is the posttest outcome score; Y_{ij0} is the baseline outcome score; $j = 1$ represents the comparison group; $j = 2$ represents the active treatment group; and X_j is a dummy code represent the treatment.

The average outcome for the comparison group is:

$$\bar{Y}_{11} = \beta_0 + \beta_1 \times 0 + \beta_2 \bar{Y}_{10}$$

The average outcome for the active treatment group is

$$\bar{Y}_{21} = \beta_0 + \beta_1 \times 1 + \beta_2 \bar{Y}_{20}$$

Subtracting the first Formula from the second Formula gives the difference between the average treatment outcome and the average comparison outcome:

$$\bar{Y}_{21} - \bar{Y}_{11} = \beta_1 + \beta_2(\bar{Y}_{20} - \bar{Y}_{10})$$

The average treatment effect in ANCOVA is assumed to be:

$$\beta_1 = (\bar{Y}_{21} - \bar{Y}_{11}) - \beta_2(\bar{Y}_{20} - \bar{Y}_{10}) \quad (3)$$

The formula (3) indicates that when there is no treatment effect, the ANCOVA null hypothesis assumes that the following equation is true: $(\bar{Y}_{21} - \bar{Y}_{11}) = \beta_2(\bar{Y}_{20} - \bar{Y}_{10})$. In other words, under ANCOVA's null hypothesis, the posttest distance between the group means will be exactly β_2 times the distance between the pretest group means, if there is no treatment effect. In addition, by comparing Formula (2) and Formula (3), we can see the difference between them is that β_2 is a fixed value of one in Formula (2) and β_2 is estimated from the within-group slope in Formula (3). Formula (3) assumes that the within-group slope is an accurate estimate of the expected shrinkage of the distance between group means from their pretest difference to their posttest difference, under the implicit null hypothesis.

Deciding whether to use the simple gain score approach or the residual gain score approach to predict change is a foundational issue in human development studies, since valid causal inferences are necessary to make applications about what people should do (instead of some alternative) to optimize change. Also, the simplest 2-wave predictions of the two types of change are the basic building blocks for all longitudinal analyses, to predict intra-individual change or within-person change. For example, the cross-lagged panel model incorporates multiple cross-lagged paths, using the residual gain score approach, whereas the usual linear slope in a latent growth model is an example of the simple gain score approach. In a two-wave latent growth model, we look at how individuals' scores change from time 1 to time 2 using the simple gain score approach.

$$\text{Level 1:} \quad Y_{ti} = \beta_{0i} + \beta_{1i}T_{ti} + \omega_{ti}$$

$$\text{Level 2:} \quad \beta_{0i} = \gamma_{00} + \gamma_{01}X_j + \mu_{0i}$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11}X_j + \mu_{1i}$$

where level 1 represents how individual scores change over time; Level 2 represents how individuals' initial scores deviate from the average initial score and how individuals' slopes (linear change) over time deviates from the average slope over time in addition to an average difference in initial scores and slopes due to the effect of X_j . In level 1, Y_{ti} represents individual i 's outcome at time t ; β_{0i} represents the starting point (when $T_{ti} = 0$) on individual i 's best-fitting line across time; and β_{1i} represents the individual's linear slope across the waves, T_{ti} . In level 2, γ_{00} represents the mean of the starting points (estimated baseline "true" scores) on the outcome when $X_j = 0$; μ_{0i} represents the deviation of the individual's baseline score from the baseline score as predicted by the fixed part of that equation; X_j is the treatment with $j = 2$ for the active treatment group and $j = 1$ for the comparison group; γ_{10} is the average slope across the waves when $X_j = 0$; and μ_{1i} is the deviation of the individual's slope from the slope predicted from the fixed effects part of that equation. The latent growth model has been used to analyze within-person changes (Level 1) as well as between-person differences in those changes (Level 2), if person-mean centering is used at Level 1. Level 1 is then the simple gain score based on the best fitting straight line for each individual's outcome scores across all waves. When applied to two waves, the slope is the simple gain score from Wave 1 to Wave 2.

Whereas the usual linear growth model uses simple gain scores as its basic building block for change, cross-lagged panel models use residual gain scores. In the cross-lagged panel model, we examine the bidirectional effects between the treatment score and the outcome score over time (Selig & Little, 2012; Shingles, 1985). The two-wave cross-lagged panel model can be described as follows:

$$X_1 = a + bX_0 + cY_0 + \varepsilon_X$$

$$Y_1 = d + eX_0 + fY_0 + \varepsilon_Y$$

Where, X and Y is the two reciprocal variables with subscripts of zero for the pretest and one for the posttest. These two formulas are a combination of two residual gain score functions for the two variables - both as predictor and outcome in different time points. Since the simple gain score approach and the residual gain score approach are fundamental building blocks for more complicated models, understanding how to make causal inferences based on these two methods will help understand the causal inference implications for more complicated model.

Lord's Paradox and Empirically Inconsistent Results

The two approaches for reducing selection bias are often not consistent with each other in nonrandomized studies, as illustrated in Lord's (1967) paradox. In Lord's hypothetical data (Figure 2), males' and females' weight gains were compared with each other using the two approaches. Males' and females' initial average weight are set to be significantly different from each other and after 9 months males end up with the same average weight as their initial average weight, as is the case for females. When using the simple gain score approach to analyze the data, since neither males nor females have gained any weight on average, the results indicated no difference in weight gained in

comparing the two gender groups. However, the residual gain approach indicated that males gained more weight than females who started at the same weight as they did. This is because male's unchanged mean weight was higher than expected from the regression of the pretest group means toward a common mean according to the null hypothesis implicit in ANCOVA. Lord's paradox reveals the existence of the inconsistency from the two approaches and indicates that at least one approach must be biased, if not both of them. It also shows that the residual gain score approach is biased in the direction of the pretest mean group *relative to* the simple gain score approach.

The inconsistency of Lord's paradox has been supported by several real data analyses. Berger and colleagues (2009) used the National Survey of Child and Adolescent Well-Being data to study the effect of out-of-home placement on child well-being with multiple statistic approaches including the simple gain score approach and the residual gain score approach. Pre-treatment differences between the stay-at-home group and the out-of-home placement group existed on SES, internalizing behavior problems, externalizing behavior problems, and the type and severity of child maltreatment. The results from the two approaches produced inconsistent results in that out-of-home placement increased both internalizing and externalizing behavior problems using the residual gain score approach (significantly and marginally, respectively), but out-of-home placement decreased externalizing problems significantly and had no effect on internalizing problems using the simple gain score approach. The differences in effects were similar to the relative differences in Lord's paradox, in that the results from the residual gain score approach were biased in the same direction as the pre-test group

means on externalizing and internalizing behavior problems, relative to the results from the simple gain score approach.

Similarly, Larzelere, Ferrer, Kuhn, and Danelia (2010) studied whether four disciplinary punishment variables and two professional interventions (therapy and Ritalin) help reduce antisocial behavior and hyperactivity in children. The residual gain approach and the simple gain score approach led to contradictory results in that all of the 12 significant results from the residual gain score approach indicated detrimental effects for all punishments and professional interventions, whereas all of the nine significant results from the simple gain score approach indicated beneficial effects for the nonphysical punishments and non-significant effects for physical punishment and the professional interventions. Again, relative to each other, results from the residual gain score approach were biased in the direction of pre-test group means of the outcome variables, compared to the simple gain score approach. These analyses indicate that Lord's paradox exists in real data analyses as well as in hypothetical data (Van Breukelen, 2013, shows examples from other data).

Understanding Implications of Lord's Paradox

Questions Raised from Lord's Paradox

Lord's paradox has triggered many discussions about which approach is unbiased and correct for making causal inferences (Allison, 1990; Holland & Rubin, 1983; Lord, 1967, 1969; Maris, 1998; Rausch, Maxwell, & Kelley, 2003; Rubin, 1974, 1977; Senn, 2006; Van Breukelen, 2006; Wainer, 1991; Wainer & Brown, 2007; Weisberg, 1979; Wright, 2006). Debates indicate that some prefer the simple gain score approach, because

the simple gain score approach may not only rule out the effect of the preexisting differences on the outcome score but also reduce some threats of spuriousness of confounds (Allison, 1990). Others criticize the simple gain score approach because the results from the simple gain score approach can be artifactual due to regression towards the mean (Campbell & Kenny, 1990; Marsh & Hau, 2002). Furthermore, when the change from the pretest to the posttest is due to natural differential growth rather than the treatment effect. The natural growth will count for treatment effect even there is no treatment effect in the simple gain score approach, and there is no way to differentiate the natural differential growth from the treatment effect (Blumberg and Porter, 1983). As a result, regression towards the mean could make the estimated treatment effect appear to be larger or smaller than what it should be in reality in the simple gain-score approach. If so, regression toward the mean could produce a significant-looking treatment effect even when there is no true treatment effect at all. Finally, simple gain scores are less reliable than the pre-test or post-test scores and therefore have less statistical power than the residual gain score approach (May & Hittner, 2010; Van Breukelen, 2013).

On the other hand, although the residual gain score approach may have more power than the simple gain score approach, when analyzing corrective actions by parents or by professional to correct problems such as children's misbehaviors, health problems, and children's homework problems, biases in the residual gain score approach may make all corrective actions look harmful (Larzelere, Lin, Payton, Washburn, in press; Larzelere & Cox, 2013). For instance, Deptula, Henry, and Schoeny's (2010) study of parents talking with their child about negative consequences of unprotected sex behaviors significantly predicted adolescent's unexpected pregnancies in the next seven years using

the residual gain score approach. The first major study of the Head Start program used the residual gain score approach and showed negative effects of Summer Head Start on child well-being (Cicirelli, 1969; Wu & Campbell, 1996). These counterintuitive results suggest biases of the residual gain score approach for analyses of corrective actions.

The debates and counterintuitive results suggest problems in these two approaches in nonrandomized studies. In contrast, Lord's paradox does not exist in randomized studies. Thus, one conclusion is that in large randomized studies, in which predictions of the residual gain score approach and the simple gain score approach are both unbiased and therefore agree with each other, the residual gain score approach is considered superior to simple gain score approach due to power considerations (Van Breukelen, 2013). The residual gain score approach has more statistical power, apparently to the extent that the pre-test explains some variance in the outcome. In addition, the residual gain score approach produces an unbiased causal estimate in the regression discontinuity design (i.e., when treatment assignments are completely due to the covariate, if its assumptions are satisfied, such as correct modeling of the linear or nonlinear regression lines), unlike the simple gain score approach (Van Breukelen, 2013). This seems to be because treatment assignment is fully accounted for by one covariate, in contrast to all other non-randomized designs. In other non-randomized designs, it is not clear that either approach is unbiased, nor is it clear which one gives a less biased causal estimate.

So why does Lord's paradox exist and what leads to the paradox in nonrandomized studies? Namely, why are the simple gain score approach and the residual gain score approach consistent and unbiased in randomized studies but inconsistent and biased in nonrandomized studies? Can we get consistent results in

nonrandomized studies? For answering these questions, it is useful to review the basic idea of making casual inferences, the assumptions of the two approaches and the consequences of violating those assumptions.

Making Valid Casual Inferences

Making a causal inference involves an interpretation of the connection between two events: A, the cause, and B, the effect. A causal relationship between A and B exists when 1) A is reliably correlated with B; 2) A happens before B happens; and 3) A is the only plausible cause of B, i.e., there is no other plausible alternative interpretation of the temporal association between A and B (Shadish et. al., 2002). Longitudinal studies can satisfy the temporal sequence requirement better than cross-sectional studies. The third requirement requires only that A is a sufficient reason to cause B. For example, spending long times watching TV is highly correlated with obesity. The reason is that when people spend a long time watching TV, they are more likely to overeat and less likely to exercise, compared to those who spend less time watching TV. Thus, the cause of obesity is overeating and/or lack of exercise but not watching TV too much. It is necessary to point it out that overeating and/or lack of exercise may not be the only reasons for obesity (e.g., some medicines can lead to gaining weight), but they are sufficient reasons to cause obesity. Namely, there are multiple paths that could lead to the same outcome.

The counterfactual model and the fundamental problem of causal inference.

The logic of the causal effect from A to B is that if A has happened, B happens; if A has not happened, B does not happen (Shadish et. al., 2002). This causal effect relationship can be explained in David Hume's counterfactual model (Lewis, 1973; Shadish et. al.,

2002; Morgan & Winship, 2007). In the counterfactual model, each studied unit has two potential outcomes: an outcome of the treatment and an outcome of the non-treated or comparison condition. If a studied unit is assigned to or self-selects to participate in an active treatment group, the unit will end up with the treatment outcome. The counterfactual is what the outcome would have been if this same studied unit had experienced the comparison condition instead of the treatment at that time. Consider a one-month treatment of psychological therapy for depression as an example. A participant will have two potential outcomes: the outcome related to psychological therapy, $Y(E)$, if he/she is treated, and the outcome related to the comparison condition, $Y(C)$, if he/she is not treated. If the participant received the therapy treatment, the counterfactual is what would have happened if the same participant had been in the comparison group instead of the treatment group. If the participant's depression symptoms decreased when he/she received the psychological therapy, and the depression symptoms would have stayed the same if he/she would not have received psychological therapy, the causal effect of psychological therapy is the difference of the participant's depression level after he/she was treated (received psychological therapy) compared to what it would have been otherwise (not received psychological therapy): $Y(E) - Y(C)$. Here, $Y(E)$ represents the treatment outcome and $Y(C)$ represents the no-treatment outcome (i.e., the counterfactual). The problem for the counterfactual model is that it is impossible to observe something happening and not happening at the same time. This is called the "fundamental problem of causal inference" (Holland, 1986).

Rubin's causal model. Although the fundamental problem makes the estimate of a causal effect impossible for any individual during the same time period, this does not

mean that making valid casual inferences is impossible. To overcome these problems, Rubin (1974), based on the concept of potential outcomes, developed his Causal Model (RCM) that extended the idea of one experimental unit of study to N experimental units being studied. In the original RCM model (1974), half of the studied units were randomly assigned to the control condition and the other half of the studied units were assigned to active treatment. Each studied group has two potential outcomes: an outcome of being treated and an outcome of not being treated. Assuming Y_1 and Y_2 represent the outcomes of two experimental groups, each individual in each group will have two potential outcomes: the treated outcome $Y(E)$ and the non-treated outcome $Y(C)$. The treatment effect for any person in Y_1 will be $[Y_{1i}(E) - Y_{1i}(C)]$ and the treatment effect for any person in Y_2 will be $[Y_{2i}(E) - Y_{2i}(C)]$. The difference between the average active treatment outcome and the average control outcome, which represents the average treatment effect Δ , will be the average treatment effect of the two groups:

$$\Delta = \frac{1}{2} \left\{ \left[\frac{\sum [Y_{1i}(E) - Y_{1i}(C)]}{N} \right] + \left[\frac{\sum [Y_{2i}(E) - Y_{2i}(C)]}{N} \right] \right\}$$

This formula could be displayed as their average score. We have:

$$\Delta = \frac{1}{2} \{ [\bar{Y}_1(E) - \bar{Y}_1(C)] + [\bar{Y}_2(E) - \bar{Y}_2(C)] \}$$

In reality, each studied group only receives one outcome, either the treated outcome or the non-treated outcome. Namely, because \bar{Y}_2 is obtained from the active treatment group and \bar{Y}_1 is obtained from the comparison group, $\bar{Y}_2(E)$ and $\bar{Y}_1(C)$ are observed, but $\bar{Y}_2(C)$ and $\bar{Y}_1(E)$ are missing. Since the two groups are randomly assigned, it is assumed that no matter which experimental group will receive which treatment in reality, the average treatment effect should be similar if they had been in the other

treatment condition. Thus, the absent outcome for one group could be estimated by the present outcome of the other group. We have the following two conditions: 1) if the two groups all received active treatment, it will have $\bar{Y}_1(E) = \bar{Y}_2(E)$; and 2) if the two groups both received the control condition, it will have $\bar{Y}_1(C) = \bar{Y}_2(C)$. These equalities justify replacing the absent $\bar{Y}_1(E)$ with the present $\bar{Y}_2(E)$ and replacing the absent $\bar{Y}_2(C)$ with the present $\bar{Y}_1(C)$, so that Equation (3) could be revised:

$$\Delta = \bar{Y}_2(E) - \bar{Y}_1(C) \quad (4)$$

where \bar{Y}_2 is the average treated outcome and \bar{Y}_1 is the average comparison outcome. In Formula (4), the foundation is that the two experimental groups may not be exactly the same, but they should not have any significant differences before getting treatment and should have a similar average response to the same treatment no matter which group receives the experimental treatment or receives the control treatment (Rosenbaum & Rubin, 1984). In this situation, the comparison group could be used to estimate the hypothesized average outcome of the active treatment group if they had not received treatment, and the estimation of the treatment effect is an “unbiased” estimate of the casual effect. This illustrates how RCM specifies the assumptions necessary for unbiased causal estimates, in this case from a randomized study.

Assumptions

The assumption of strong ignorability. The RCM includes a fundamental assumption, the strongly ignorable treatment assignment assumption (Rubin, 1978). When using Rubin’s Causal Model in randomized studies, since the two experimental groups are similar, it is assumed that no matter which experimental group will receive

which treatment condition in reality, the treatment effect should be similar if they reversed treatment assignments. Namely, if the experimental units in the comparison group received the same treatment as the active treatment group and the treatment group did not, the treatment effect for the comparison group should be the same as the original active treatment group. For example, 60 classes are randomly assigned to two experimental groups: A and B. The treatment effect, that if group A is assigned to a special writing program and group B is assigned to the traditional writing program, should be the same as the treatment effect if group B had been assigned to the special writing program and group A had been assigned to the traditional writing program. In this situation, how to assign the study units will not influence the treatment effect. Thus, the treatment assignment itself could be ignored since which group is assigned to active treatment or comparison treatment is expected to be unrelated to the treatment effect, due to random assignment. The fundamental assumption of strongly ignorable treatment assignment is that the two experimental groups should be similar or have matched prognoses, which could be satisfied in successfully randomized designs. Thus, the randomized research design should be viewed as an unbiased design. In nonrandomized studies, when pretest differences exist between the comparison group and the treatment group and the two experimental groups are not matched on what their prognoses would be other than treatment, the result is biased because the comparison group could not be used to estimate the absent non-treated effect in the treatment group.

The assumption of strong ignorable treatment assignment is essential for making valid casual inferences, which must be satisfied to use either the simple gain score approach or the residual gain score approach for making unbiased causal inferences. The

application of the simple gain score approach and the residual gain score approach for “controlling for” pre-treatment difference in nonrandomized studies is a misunderstanding of using these two approaches because they were not originally designed for “controlling for” pre-treatment difference. Rubin (1974) and Campbell (Shadish, 2010) both agree that when the causal effect could be explained by third confounding variables other than the treatment effect, we have to control for the effect of the confounding variables. However, controlling a confounding effect is most rigorously based on randomized studies under the assumption that the pre-treatment conditions between the comparison group and the active treatment group are matched and comparable. When Rubin (1974) introduced his Causal Model using the simple gain score approach to control for additional variables, such as different pretest scores, he said that the adjusted scores are unbiased “given random assignment” (p. 696). He emphasized that the unbiased estimation otherwise needs to assume that “some ‘known’ function for x_j (e.g., in the compensatory reading program example, suppose x_j equals $[.10 \times IQ] \times pretest \times [percentile\ of\ family\ income]$), so that x_j is the same whether the j^{th} unit received E or C” (p. 696). Moreover, although Campbell and Stanley (1963) prefer the ANCOVA approach for making casual inferences, they were careful to state that “the usual statistics (including ANCOVA, cited earlier on the same page) are appropriate only where individual students have been assigned at random to treatments” (p. 23). In non-randomized designs, it is usually misleading to claim valid causal inferences when using either the simple gain score approach (May & Hittner, 2010; van Breukelen, 2013) or the residual gain score approach (Miller & Chapman, 2001) to

“adjust” the pretest difference because neither “adjustment” satisfies the assumption of strong ignorable treatment assignment..

The assumptions of ANCOVA. In nonrandomized studies, violations of assumptions not only involve the assumption of strong ignorability, but also involve other assumptions of the residual gain score approach and the simple gain score approach such as normality of outcome residuals, homogeneity of slope, and homogeneity of variance, which could lead to biased results. ANCOVA requires the following assumptions: 1) normality of the dependent variable’s residuals; 2) homogenous variances in all the compared groups; 3) independence of each participant’s scores (Rausch, Maxwell & Kelley, 2003); 4) linearity of regression; 5) homogeneity of slopes (Keppel & Wickens, 2004); 6) no measurement error in the covariate; and 7) independence between covariates and treatment (Huitema, 2011; Van Breukelen, 2013). The first four assumptions are basic assumptions for linear regression. The last three assumptions are special assumptions for ANCOVA. When the assumption of independence of treatment and covariate is violated, violation of homogeneous slopes will also increase the type I error rate (Hamilton, 1976; Hollingsworth, 1980; Huitema, 2011; Rogosa, 1980). Furthermore, the combination of violated assumptions of homogeneous slopes, equal group sample sizes, and normality can distort the results of ANCOVA (Levy, 1980; Sullivan, & D’Agostino, 2002).

The last assumption of independence between treatment assignment and covariates means no covariates difference, e.g., no mean difference on pretest score on the outcome variable, between the active treatment group and the comparison group. The original main purpose of ANCOVA was to reduce the standard error by eliminating part

of the extraneous influences on the outcome variable, NOT to adjust for pre-treatment differences on a covariate (Miller & Chapman, 2001). The assumption of independence between treatment and covariates emphasizes that each covariate should have a population mean common to all groups, which means that the covariates do not have group mean differences at the population level. Then regression towards this common covariate mean could be used to estimate the posttest mean difference (Huitema, 2011). This could be satisfied in perfect randomized studies since the two experimental groups are matched on pretreatment conditions. However, in nonrandomized studies, the covariate is likely to be have different pre-test means in the experimental groups, indicating no common covariate mean, in which case the covariate adjustment will be biased (Huitema, 2011).

What would occur if treatment assignment depends on the covariate? Recall the adjusted average treatment effect using the ANCOVA approach in Formula (3):

$$\beta_1 = (\bar{Y}_{21} - \bar{Y}_{11}) - \beta_2(\bar{Y}_{20} - \bar{Y}_{10}) \quad (3)$$

The estimated average treatment effect depends on the posttest mean difference, the pretest mean difference, and the slope β_2 predicting the effect of pretest on posttest. When there is no pretest mean difference between the two experimental groups, $\bar{Y}_{20} - \bar{Y}_{10} = 0$, we have $\beta_1 = (\bar{Y}_{21} - \bar{Y}_{11})$, which means the average treatment effect is the only contribution for the posttest mean difference or the between-person difference. In contrast, when the covariate group means actually differ from the grand mean, the average treatment effect is NOT the only explanation for the posttest mean difference but it is also affected by the pretest difference. This has been illustrated by Huitema (2011)

and Van Breukelen (2013). In Figure 1, Δ_{xr} represents the estimated posttest difference when there is no pretest difference and $\bar{Y}_{21} - \bar{Y}_{11}$ represents the estimated posttest difference where there is a pretest difference. When there is no pretest difference, the posttest difference Δ_{xr} is the explanation of the treatment effect since no other factors make the posttest different between the treatment group and the comparison group. However, if the pretest means are different, the posttest difference is $\bar{Y}_{21} - \bar{Y}_{11}$, which includes not only the treatment effect but also the effect attributed to the pretest mean difference. In this situation, treatment is not the only reason that makes the posttest difference and an estimation of posttest difference for estimating treatment effect may be biased. To get unbiased results that $\Delta_{xr} = \bar{Y}_{21} - \bar{Y}_{11}$, the pretest group means \bar{Y}_{10} and \bar{Y}_{20} need to be equal to the grand mean \bar{Y}_0 .

To further understand the consequence of violating the assumption of independence between treatment assignment and covariates, Formula (2), the ANCOVA formula, is modified as follows:

$$Y_1 = \beta_0 + \beta_1 X + \beta_2 Y_0 + \varepsilon \quad (5)$$

Where, Y_1 is the posttest outcome score; Y_0 is the baseline outcome score; and X is a dummy variable to identify the comparison group (0) or the active treatment group (1).

Applying differential calculus shows the change of Y_1 :

$$\Delta Y_1 = \beta_1 \times \Delta X + \beta_2 \times \Delta Y_0$$

Change in Y due to change in X is estimated by the first derivative with respect to X :

$$\frac{\Delta Y_1}{\Delta X} = \beta_1 + \beta_2 \frac{\Delta Y_0}{\Delta X}$$

The purpose of converting the original ANCOVA formula to its solution according to differential calculus is to detect factors that relate to the change of outcome due to being treated or not being treated (Brorsen, personal communication, March 9, 2018). The term β_1 is ordinarily interpreted as the average treatment effect; ΔY_1 is the change in the posttest outcome associated with ΔX , the change from control to treatment condition; ΔY_0 is the change in the pretest scores predicted by the change from control to treatment condition. In the equation, $\frac{\Delta Y_1}{\Delta X}$ estimates the amount of the posttest change associated with one unit change of treatment condition, indicating how posttest change depends on treatment change, which is the observed treatment effect in unadjusted posttest outcomes; $\frac{\Delta Y_0}{\Delta X}$ estimates the amount of pretest change associated with one unit change of treatment condition, indicating how pretest differences are associated with treatment assignment. ANCOVA usually interprets β_1 as the treatment effect, but that is an unbiased estimate of the treatment effect if the change of posttest outcome is only influenced by the change of treatment condition (the treatment assignment). To satisfy this, it requires $\frac{\Delta Y_1}{\Delta X} = \beta_1$, which means $\frac{\Delta Y_0}{\Delta X} = 0$, indicating treatment assignment is independent of the pretest outcome. If $\frac{\Delta Y_0}{\Delta X} \neq 0$, the changes from the pretest outcome to the posttest outcome will not only be dependent on the treatment condition but also will depend on the relationship between the pretest score and the treatment assignment, indicating the estimated treatment effect β_1 is biased. Since $\beta_2 \frac{\Delta Y_0}{\Delta X}$ could be negative or positive, the biased results could be in either direction.

In sum, it is a misunderstanding that ANCOVA is guaranteed to “control for” pre-treatment mean differences and produce unbiased results in nonrandomized studies.

Unfortunately, human development investigators frequently favor the residual gain score approach to “control for” or to “remove” differences between the comparison group and the active treatment groups in non-randomized studies. When this mistake is pointed out, the investigators’ reactions are usually surprised (Miller, 2001). Furthermore, in nonrandomized studies, violations of assumptions such as homogeneity of slope and homogeneity of variance could lead to biased results. However, it is common for human developmental investigators to ignore these assumption tests, even though they may cause biases in the results. The present study will use simulated data to test the consequence of violating some of these assumptions (normality, homogeneous variances, homogeneous slopes, and independence between treatment and covarates). I also test assumptions of the simple gain score approach.

The assumption of the simple gain score approach. Similar to the residual gain score approach, the simple gain score approach relies on assumptions from repeated measures ANOVA: 1) continuous and normally distributed outcome residuals, 2) no significant outliers, 3) sphericity, i.e., that all the variances and covariances of the difference scores among all combinations of occasions must be equal, and 4) homogeneity of covariance (Keppel & Wickens, 2004). Compared to the residual gain score approach, the simple gain score approach does not holds unique assumptions except the sphericity assumption. The sphericity assumption is a special assumption for repeated measures ANOVA (i.e., the simple gain score approach or CHANGE). A sphericity condition is satisfied when the “variances of difference scores for all pairs of treatments” (p. 270, Maxwell, 1980) are homogeneous (Huynh, 1978). Violation of these assumptions could also lead to biased results (Huitema, 2011). With two waves, since there is only one

within-person variance of difference scores, the sphericity is automatically satisfied. Thus, the sphericity assumption will not be tested in this study. The fourth assumption generalizes the homogeneity of variance to require that the correlation of pre- and post-test scores are similar across groups. This additional assumption beyond homogeneity of variance seems to be equivalent to the homogeneity of slopes in ANCOVA. This last assumption applies to the between-participant results (e.g., treatment group vs. comparison group), whereas the sphericity assumption refers to the within-participant results.

In addition to assumption violations, how could the simple gain score approach produce biased results? To show the biased results of the simple gain score approach, I modify Formula (1), the simple gain approach, as follows:

$$Y_1 - Y_0 = \alpha_1 + \alpha_2 X + e$$

Here, Y_0 is the pretest score; Y_1 is the posttest score; and X is a dummy variable to identify the comparison group (0) or the treatment group (1). By subtracting Y_0 from both sides of the equal sign, we get:

$$Y_1 = \alpha_1 + \alpha_2 X + Y_0 + e \tag{6}$$

Formula (6) indicates that the coefficient for the pretest Y_0 is one, indicating that each individual's estimated posttest score under the null hypothesis is identical to the pretest score (plus an overall intercept of α_1), which is impossible in standardized scores, if there is any time-related variability in rank order across the two waves, even if due only to measurement error. This is impossible, however, only given the usual ANCOVA assumption that the within-group slope coefficient is used to estimate how much the pretest group means will come closer together under the null hypothesis. In Lord's (1967)

paradox, the coefficient is 1.00 for estimating shrinkage of the group mean differences in Y_0 from pretest to posttest, even though the within-group slope is substantially less than 1.00, estimated herein to be .48.

If, however, we retain the ANCOVA assumption that shrinkage of the distance between group means from the pretest to the posttest is estimated by the within-group slope, under the null hypothesis, the correct formula is:

$$Y_1 - Y_0 = \alpha_1 + \alpha_2 X + \gamma Y_0 + e \quad (7)$$

where, γ would be less than zero, unless the variance is larger for Y_1 than for Y_0 , which makes it possible for γ to be greater than zero. After standardizing Y_1 and Y_0 on the basis of their respective *SDs*, γ must be less than zero. Ignoring the estimation of γ by setting it to zero will usually lead to an overestimation of the treatment effect for corrective actions. Formula (7) combines the equations for both simple gain score and residual gain score.

When we move Y_0 to the right side of the equation, we have the estimated parameter for Y_0 is $(1+\gamma)$, where $\gamma \leq 0$ if the *SD* of Y remains constant. Although $(1+\gamma)$ in Formula (7) will be less than one, this does not mean that $(1 + \gamma)$ is correctly estimated in the ANCOVA formula. The problem for ANCOVA is that the within-group slope β_2 in Formula (2) may not be the correct estimate of the $(1 + \gamma)$ to be used to indicate regression of the pretest group means toward a common mean according to the null hypothesis, when the treatment assignment are affected by the covariate, as I mentioned before.

So in what conditions can we get consistent results from the simple gain score approach and the residual gain score approach? Compared to the equation for predicting simple gain scores (Formula 6), Equation (5) shows that the treatment effect in ANCOVA is identical to the treatment effect in simple gain scores only when $\beta_2=1$, which rarely happens in reality. But that is impossible only under the standard ANCOVA assumption that equates the within-group slope with the value used to estimate shrinkage of the distance between group means on Y from the pretest to the posttest. Another possibility to reach consistent results for the two approaches is to have equal pretest group means, which is exactly illustrated as ANCOVA's assumption of independence between the treatment conditions and the covariate.

Another Issue: Under-Adjustment

If analyses from the simple gain score approach and the residual gain score approach provide consistent results, does it prove that the results are unbiased? Rubin (1974) emphasized that when using the simple gain score approach, 1) whether we can get unbiased results depends on whether the model is appropriate or not; 2) we may never know all the confounding effects in reality; and 3) if we satisfy the strongly ignorable assumption, causal inference are unbiased. The strong ignobility assumption is an assumption necessary for unbiased causal estimates, but it is usually difficult to accomplish and impossible to test, at least completely. There are somewhat equivalent (or overlapping) assumptions that are essential for valid causal inferences that signify that, if our statistical model is a perfect representation of reality, then our causal estimates are unbiased. But our statistical models are never perfect representations of reality. For instance, in human development studies, most research designs are nonrandomized

studies, which are more complicated so that known or unknown confounds could lead to pretest differences between the comparison group and the active treatment group, compared to randomized studies. Even if either including the pretest outcome as a covariate or simply subtracting the pretest from the posttest outcome could reduce selection bias, it doesn't truly reveal the mechanism through which variables will lead to selection bias (Cook, Shadish, & Wong, 2008). In other words, our statistical model is not a perfect representation of reality. Imperfection in partialing out important confounds has been called the under-adjustment bias (Campbell & Boruch, 1975). Thus, the key problem leading to biased results is the existence of unknown pre-treatment differences between the active treatment group and the comparison group that could cause differences in the two groups' outcomes (Van Breukelen, 2013). Then, the problem becomes how can we approximate the required perfection and how can we tell whether we have approximated that well enough.

What will be effective to balance the preexisting difference between the comparison group and the active treatment group? The consensus among most researchers is that under the condition that selection bias cannot be truly balanced, it is better to include the pretest score on the outcome since the pretest is the best predictor for the posttest outcome and it may control for the effect of other confounds (Diaz & Handa, 2006). Also, some researchers suggest including all measured variables as predictors to minimize the issue of under-adjustment in controls.

In sum, in nonrandomized studies, due to the unmatched nature of the comparison group and the active treatment group, the results will usually be biased and could lead to inappropriate interpretations of the causal relationship between the treatment and the

effect when using the simple gain score approach or the residual gain score approach. However, this problem has rarely been raised in human development studies. In addition, although an effective method to detect which approach is less biased may not be feasible, the analysis of applying both simple gain score and residual gain score should be conducted to investigate whether we can get consistent results across both of them (Duncan, Engel, Claessens, & Dowsett, 2014) when testing these models. However, to my knowledge, there are few studies that systematically compare the consistency of causal estimates across alternative longitudinal analyses. Questions remain whether results from modified models that satisfy the assumptions can be consistent across analyses of the residual gain score approach and the simple gain score approach. Moreover, assuming a lack of consistency across analyses, what models and methods can either (1) improve the validity of causal inferences and/or (2) produce the consistency across analyses that would occur in idealized randomized studies? The answers have not been not clear.

Possible Solutions

To answer these questions, the first step is to test whether violations of the assumptions of normality, homogenous slopes, homogenous variances, and homogeneity of covariate means contribute to Lord's (1967) paradox. If the effect of violating a certain assumption is trivial, model modification will not be considered to address that assumption. Otherwise, alternative models will be considered for satisfying the assumption or for dealing more appropriately with violating it. I assume that discrepancies in means, or slopes, or variances are more likely to occur when (1) there are distinct subgroups that vary on their prognosis on the outcome and (2) the most at-

risk groups are over-represented in the treatment condition compared to the control condition. By identifying homogeneous trajectory subgroups, comparisons between treatment conditions can then be closer to comparing otherwise equivalent groups than when heterogeneous subgroups are analyzed together as though they constituted one homogeneous group. If the active treatment group and the comparison group are from one homogenous group, their variances, slopes, and covariate means will be closer to each other. The important issue is to have similar groups in each treatment condition, in which case each group's covariates means will be similar to each other. How can we get similar groups?

The Latent Class Growth Model

The latent class growth model, a special type of mixture model that uses the repeated measures of longitudinal outcome variables to identify heterogeneous classes/groups of growth trajectories, is first considered. The idea of the latent class growth model is to maximize between group differences of subgroups that have similar within-person patterns/trajectories over time (Curran, Obeidat, & Losardo, 2010). The changing trajectories for all the individuals within each subgroup are similar and the subgroups have more homogeneous pretest scores on the outcome variable than the entire group (Jung & Wickrama, 2008). The latent class model could be used to identify homogeneous trajectory subgroups and conduct analyses within each homogeneous trajectory subgroup. Within each trajectory subgroup, the effect of other covariates that could influence the posttest outcome may or may not be balanced out (Haviland, Nagin, & Rosenbaum, 2007). It is expected that within each subgroup, the variances, the slopes,

and pretest means between the active treatment group and the comparison group will be closer to each other than in the original heterogeneous group.

If within each homogeneous trajectory subgroup, the variances, the slopes, and pretest means between the active treatment group and the comparison group are closer than the heterogeneous subgroups as a whole but consistent results are not achieved within a homogeneous trajectory subgroup, then removing pretest differences to fit the assumption of ANCOVA that no pretest differences effect on treatment will be considered. ANCOVA assumes that a zero treatment effect will be shown when the group means regress toward the grand mean to the extent predicted by the within-group autoregressive correlations between the pre-test and the post-test. That seems quite reasonable when the groups being compared are the same, including having the same mean and autoregressive within-group correlation. However, it violates the assumption of ANCOVA when the groups differ. The solution for satisfying the ANCOVA assumption of independence between covariate and treatment needs to focus on removing the pretest difference between the treatment group and the comparison group.

The Group-Centered ANCOVA

To remove the pretest difference, I recommend a modified *group-centered ANCOVA*, adapted from Huitema's (2011) quasi-ANCOVA approach originally designed for analyzing a treatment effect when a covariate is measured after treatment. The group-centered ANCOVA approach is a modification of ANCOVA by centering both pre-test and post-test scores on the group pretest means. Through centering the pretest outcome around the pretest group means, the group difference in pretest means would be zero. By

centering the posttest outcome score around the pretest group mean, the relative difference in simple gain scores remains the same between the active treatment and comparison groups. The comparison of the ANCOVA approach and the group-centered ANCOVA approach is shown in Formula (8) and Formula (9). In the traditional ANCOVA approach, the individual score would be centered around the grand mean and presented as follows:

$$Y_{1ij} = \beta_0 + \beta_1 X_{ij} + \beta_2 (Y_{0ij} - \bar{Y}_0) + \varepsilon \quad (8)$$

where Y_{1i} is the posttest score of the i^{th} individual in the j^{th} group; Y_{0ij} is the pretest score of the i^{th} individual in the j^{th} group; X_{ij} is a dummy code to identify the treatment; and \bar{Y}_0 is the grand mean. When using group-centered ANCOVA, the formula (8) could be modified as following:

$$Y_{1ij} - \bar{Y}_{0j} = \delta_0 + \delta_1 X_{ij} + \delta_2 (Y_{0ij} - \bar{Y}_{0j}) + \varepsilon \quad (9)$$

where Y_{1ij} is the posttest score of the i^{th} individual in the j^{th} group; Y_{0ij} is the pretest score of the i^{th} individual in the j^{th} group; X_{ij} is a dummy code to identify the treatment; \bar{Y}_{0j} is the pretest mean for group j . By centering both the pretest score and the posttest score on the pretest group mean, participants retain exactly the same increase or decrease on Y that they had in their original scores on Y , but the group-mean centered pretest scores now satisfy the ANCOVA assumption of equal group means on the covariate. The group-centered ANCOVA approach has an additional advantage. It is easy to use for the analysis when the treatment is a continuous variable by treating the continuous treatment as multiple treatment groups and centering the pretest outcome and

the posttest outcome around the pretest mean score for each score on the treatment variable. It is also easier to run compared to the mixture modeling method or the propensity score approach, which are used to minimize selection bias by generating homogeneous subgroups on the pretest outcome trajectory or by matching samples of the active treatment group and the comparison group based on propensity scores. However, the limitation for the group-centered ANCOVA approach is that the pretest mean scores are not actually equal for the active treatment group and the comparison group in reality and we do not know how much their mean difference would have changed under the null hypothesis. Thus, although the mean pretest difference between the active treatment group and comparison group is statistically removed to meet the assumption of ANCOVA, the results could remain biased in nonrandomized studies due to a true interdependent relationship between the treatment and the covariate, which has only been artificially removed by centering.

The Propensity Score Adjustment Approach

In recent years, the propensity score adjustment approach (Rosenbaum & Rubin, 1983), matching samples of the active treatment group and the comparison group based on propensity scores, has been gaining attention. Instead of adjusting for only the mean pretest difference between the treatment and the control groups, the propensity score approach imitates some characteristics of randomized studies to apply to non-randomized studies for making causal inferences. As aforementioned about the assumption of the strong ignorable treatment assignment, in perfectly randomized studies, the active treatment group and the comparison group have similar pretest characteristics on average on all variables due to being randomly assigned to the two groups. Thus, the absent

outcome for one group could be estimated by the present outcome of the other group and the average treatment effect could be estimated without bias by accounting for the present treatment outcome and ignoring the absent treatment outcome. However, in nonrandomized studies, since the active treatment group and control treatment group have different pretest characteristics, the absent outcome for one group could not be estimated by the present outcome of the other group. In order to overcome this problem, the propensity score approach adapted the idea of balancing samples between the active treatment group and the comparison group in randomized studies, employing propensity scores to balance samples between the active treatment group and the comparison group in non-randomized studies. Here “the propensity score, is a balancing score: conditional on the propensity score, the distribution of measured baseline covariates is similar between treated and untreated subjects” (Austin, 2011). The propensity score is the probability of a particular case to be in a treatment group based on all the covariates. Ideally, samples with the same propensity score in the two groups will have similar characteristics on all the baseline confounding variables. Thus, conditional on the propensity score, the samples with the same propensity score will have an equal chance to be in the active treatment group or in the comparison group. If two participants in the two experimental groups have the same propensity scores, it is assumed that the absent control condition of a participant in the treatment group could be estimated by the present outcome of the other participant in the comparison group if there are no unknown effects other than the covariates’ effects on the outcome. In nonrandomized studies, since the true propensity score is unknown due to unmeasured effects, and there is no guarantee

that all pretest conditions are balanced, checking the sample balance on the covariates is necessary.

Using the propensity score method, the first step is to identify covariates for calculating propensity score for each sample. Generally, four sets of covariates are considered to be included in the propensity score model: “all baseline covariates, all baseline covariates that are associated with treatment assignment, all covariates that affect the outcome (i.e., the potential confounders), and all covariates that affect both treatment assignment and the outcome (i.e., the true confounders)” (Austin, 2011, p. 414). According to Austin (2011) and Austin and colleagues (2007), either all covariates influencing the outcome (the potential confounders) or all covariates influencing both the outcome and the treatment assignment (the true confounders) have more merit than the other two. Once covariates are selected, they are used to calculate the propensity score by applying a simple logistic regression. Then propensity scores are used to match samples to get balanced samples in the two experimental groups. If matching is successful, the matched data will be used to test whether results from the simple gain score approach and the residual gain score are consistent or not.

Current Study

The current study investigated the problems that led to Lord’s (1967) paradox using simulated data and systematically compare several models for making causal inferences in non-randomized longitudinal studies. Then it compared selected models using data from the Fragile Family and Child Well-being Study (FFCW) on treatments for depression in mothers. First, I compared the simplest applications of the two

traditional methods, the simple gain score approach and the residual gain score approach, to investigate whether the two approaches led to contradictory results, similar to Lord's (1967) paradox. I used simulated data to match Lord's (1967) paradox in a two-wave analysis, and used the FFCW data to compare the results in two-wave and in three-wave analyses. The cross-lagged panel model and a linear growth model were implemented across three waves, which incorporate the two traditional adjustment methods, respectively. The aim of the first analyses were to verify the expected contradictory results from the two common types of analyses. Then I diagnosed whether the data meet the assumptions for analyses of simple gain scores and ANCOVA and then adjust those models to satisfy those assumptions, when possible. Finally, I evaluated the statistical models on 1) which ones produce consistent results for both types of change scores and 2) any evidence or basis for thinking that some models minimize systematic bias more than other and therefore produce better approximation of unbiased causal estimates. The following models were compared: latent class growth modeling, group-centered ANCOVA, and a propensity score method. I proposed four research questions and several hypotheses (Table 1).

Research Question 1: Can I get consistent results from the two traditional methods (simple gain score approach and residual gain score approach) in two-wave analyses without other covariates, using simulated data and the FFCW data on treatment for depression in mothers? The hypothesis is that the results from the simple gain score approach and the residual gain score approach will be contradictory from the FFCW data on treatments for depression in mothers as well as the simulation data.

Research Question 2: Can I get consistent results from the cross-lagged panel model and the latent growth model from a three-wave data without any covariates, using the FFCW data on treatments for depression in mothers? I hypothesize that consistent results will not be achieved from these two approaches. Our preliminary results found significantly harmful cross-lagged path coefficients for treatments for depression, using a measure that only differentiated three levels of depression severity (Larzelere, Washburn, Lin, & Cox, 2017). The dissertation will determine whether this replicates with a more continuous measure of depressive symptom severity.

Research Question 3: What simple gain score and ANCOVA assumptions might account at least partly for the expected discrepancy between analyses predicting residual change scores vs. simple change scores? Five assumptions will be evaluated: a normally distributed outcome variable, equality of variance between the two groups, equivalent within-group slopes from the covariate (pre-test at the 1st Wave) to the outcome (post-test on the next Wave), equivalent group mean scores on the covariate (pre-test at 1st Wave), and a homogeneous pretest depression trajectory group rather than a heterogeneous mixture of pretest depression subgroups. The first two are assumptions common to the simple gain score approach and ANCOVA. The rest of them are assumptions of ANCOVA. The assumptions will be diagnosed one at a time and in combination to determine whether violations of these assumptions account for some or all of the expected discrepancies between analyses of simple gain scores vs. residual gain scores. The overall research question is whether I can get consistent results of the simple gain score approach and the residual gain score approach for the FFCW data after data

diagnostics for meeting several assumptions. Specific assumption diagnostics are the following:

3.1. Are the outcome variables normally distributed? I assume the depression variables are not normally distributed, but are consistent with a zero-inflated continuous distribution. The most appropriate link function will be used to determine whether that adjustment reduces the expected inconsistency between analyses of residual vs. simple gain scores in the FFCW data. A similar violation of the normality assumption will also be simulated.

3.2. Is the variance the same for the active treatment group and the comparison group? I assume there will be heterogeneous variance between the active treatment group and the comparison group for the FFCW data. The violation of equality of variance will also be investigated in the simulated data.

3.3. Is the slope the same for the active treatment group and the comparison group? I assume there will be nonparallel slopes between the active treatment group and the comparison group for the FFCW data. The violation of parallel slopes will also be simulated.

3.4. Also, are the covariate (pretest) mean scores on the outcome variable the same across the treatment and control subgroups (or the male and female groups in the simulated data)? I assume that the pretest means will differ for the active treatment group and the comparison group. If, as expected, the groups differ on pretest means, then that could have two consequences: First, the usual estimate for the slope from pre-test to post-test may be incorrect, even if the within-group slopes are equal to each other. If so, this

distortion of the slope in standard ANCOVA would be greater for larger differences in the pretest means of the groups. Second, analyses predicting residual gain scores may have the incorrect regression coefficient for predicting how much the group means will regress toward the grand mean from pre-test to post-test under the null hypothesis. The data for Lord's paradox are set up so that there is no regression toward the mean of the group means from pre-test to post-test (van Breukelen, 2013, equation (2), pp. 901-902). In contrast, real data, such as the FFCW data probably have some regression toward the grand mean, which could reflect spontaneous regression toward the mean expected under the null hypothesis or it could reflect a treatment effect. At this point, it is impossible to distinguish between the two explanations for the group means moving closer on their post-test means than on their pre-test means. At this step, the hypothesis is that the group means for active treatment and comparison groups from the FFCW data will be closer on the post-test than on the pre-test, in contrast to the data simulated to represent Lord's paradox. Violation of the assumption of independence between treatment group and the comparison group will be varied in the simulated data.

3.5. If the participants in the FFCW data do not represent one homogeneous trajectory group on developmental trends in depression severity before getting treatment, how many sub-groups are there? I assume the participants could be sorted into multiple sub-groups, such as the following three: a high depression sub-group, a low depression sub-group, and a no depression symptom sub-group. Mixture modeling will be used to explore this possibility.

3.6. Combine several modifications of the standard analyses based on the diagnoses of ANCOVA assumptions from 3.1 to 3.5. One possible result is that the

depression variables fit a zero-inflated continuous distribution (with non-zero scores robust to their deviations from a normally distributed, due to the Central Limit Theorem), that within-group slopes and mean covariate (pre-test) scores are different between the active treatment group and the comparison group, and that the depression developmental trends indicate multiple sub-groups.

How could model adjustments improve the results, based on these assumption diagnostics? I assume that if ANCOVA-like analyses predicting residual gain scores are adjusted appropriately when one or more of its standard assumptions are not met, then the results for predicting residual gain scores vs. simple gain scores will be closer together than when violating the standard ANCOVA assumptions. If so, which corrections are most useful for bringing analyses of residual vs. simple gain scores closer to each other? Corrections are clear for violations of these assumptions except for the assumption of equal pretest means for the treatment and control groups. First, the combination of logistic regression and multiple linear regression, called the hurdle model, within each subgroup from the mixture modeling may improve model fit, but I assume that I cannot get consistent results from the simple gain score approach and the residual gain score approach by using a combination of logistic regression and multiple linear regression within the homogeneous trajectory subgroups. Second, the group-centered ANCOVA approach will be used to recode the data by centering the pretest score and the posttest score with the pretest group mean. Because this centering will remove the group difference between their two pretest outcome means, I assume that the results from the two approaches will be consistent, but they could still be biased. If so, how biased would the results be? Then the results from the group-centered ANCOVA approach within each

subgroup will be compared to Haviland and colleagues' (2007) approach that used a propensity score approach within each subgroup.

Research Question 4: The final research question investigates whether the combinations of two methods used by Haviland et al. (2007) would produce consistent results that could be considered less biased causal estimates. They obtained consistent results across both types of gain scores by combining a type of latent class growth modeling with propensity-score methods. The latent class growth modeling balances the pretest outcome trajectory and the propensity score method balances other relative covariates. Considering this successful study and the results of the diagnostics for relevant assumptions, Research Question 4 is whether the consistency of the results can be improved further or the remaining bias be reduced by combining propensity score adjustments with latent class growth modeling as in Haviland and colleagues (2007). They also only investigated treatment conditions that started between specified waves. To follow that part of their example, the analyses dropped those who had either treatment for depression prior to Wave 4 and only compare those who started either or both types of treatment for depression between Wave 3 and Wave 4 (to the extent we can tell that from the available data). Propensity scores and homogeneous trajectory groups were based on data up through Wave 3. It is hypothesized that I can get consistent results from the two baseline adjustment methods when the adjustment is based on the propensity scores in homogeneous trajectory subgroups, while using logistic regression plus multiple linear regression. The analyses were tested separately and in combination. This may provide clues as to whether the consistent results I get from one method (e.g., group-centered ANCOVA) change very much when I use another method (e.g., Haviland and colleagues,

combined methods). If the results change much when I add methods known to improve causal estimates, that would suggest that Haviland et al.'s (2007) methods reduce the remaining bias more than group-centered ANCOVA, even if both results produce equivalent consistency.

These four research questions were tested in the next two chapters: Study I used the simulated data to test Research Questions 1, 3, and 4. The following chapter then used the FFCW data to test all four research questions.

CHAPTER III

STUDY I: SIMULATION

The purpose of the simulation is to identify problems associated with Lord's (1967) paradox. First, I start by simulating the data for Lord's (1967) paradox, which can be construed as data simulated to fit the null hypothesis according to simple gain scores, i.e., that the simple gain scores in weight were identical for males and females. For comparison, I also simulated what I am calling reversed Lord's (1967) paradox, by simulating data to fit the null hypothesis according to ANCOVA. Whereas others have debated which analysis is correct, these two simulations allow each analysis to be correct in one simulation. It was assumed that analyses of the two types of gain scores would contradict each other in the first simulation, but less has been known about the expected results from simulating reversed Lord's paradox. The subsequent analyses were then designed to diagnose the problems by testing the assumptions of the simple gain score approach and ANCOVA to determine which assumptions were associated with the expected contradictory results across the two approaches. Assumptions were tested include the normal distribution of the outcome residuals, one homogeneous group vs. a mixture of multiple subgroups, homogeneous slopes, homogeneous variances, and ANCOVA's assumption that treatment and covariate are independent.

Methods

Testing Research Question 1: Simulating Lord's Paradox and Reversed Lord's Paradox

By mimicking Lord's (1967) paradox (Figure 2), the simulated data generated 1000 students (50% females) to compare gender differences in weight gained during one year. The syntax is given in Appendix A (R Syntax for Simulating Lord's Paradox and Violated Assumptions). According to some literature (Butler, Black, Blue, Gretebeck, 2004; Filla, Hays, Gonzales, & Hakkak, 2013), the average weight of college students ranges from 140 to 145 pounds, standard deviations range from 10-25 pounds, and males' average weight is around 15 pounds above the overall mean weight. For the current simulation data, the pretest and posttest weights were set to be: mean and standard deviation for females in the pretest and posttest ($M = 130$, $SD = 15$), and mean and standard deviation for males in the pretest and posttest ($M = 160$, $SD = 15$). Note that both males and females have different mean weights but the same standard deviation. By measuring the angle between the x-axis and the linear slope in Lord's (1967) figure (about 25.5°), the slope was estimated to be around 0.477. Thus, the correlation between pretest and posttest was fixed to be 0.48 for both males and females. One way to look at these data is that they are specified to fit the null hypothesis according to the simple gain score approach as in repeated measures ANOVA, although Lord's paradox has rarely been understood from that perspective.

For comparison, I simulated a reversed version of Lord's paradox, by setting the means and SD s to fit the null hypothesis according to the residual gain score approach. To

accomplish that, I used the same pre-test means and *SDs* ($M = 130$, $SD = 15$) for females and ($M = 160$, $SD = 15$) for males, but changed the post-test means to 152.2 for males and 137.8 for females, keeping the within-gender *SDs* at 15. For simulating reversed Lord's paradox, I used the ANCOVA Formula (2) to calculate parameters for the error and the intercept, assuming the null hypothesis of ANCOVA that there is no treatment effect ($\beta_1 = 0$) is correct. When the pretest and the posttest grand means are both set to be 145, we have the intercept:

$$\beta_0 = \bar{Y}_1 - \beta_2 \times \bar{Y}_0 = 145 - 0.48 \times 145 = 75.4$$

Since the variance of Y_1 is:

$$var(Y_1) = var(\beta_2 Y_0) + var(e_{ij}) = \beta_2^2 var(Y_0) + var(e_{ij})$$

So we have,

$$var(e_{ij}) = var(Y_1) - \beta_2^2 \times var(Y_0) = 15^2 - 0.48^2 \times 15^2,$$

Then, $var(e_{ij}) = 173.16 = 13.159027^2$. $SD(e_{ij})$ is the square root of its variance, which gives 13.159. The post-test means were generated by regressing the pre-test means toward the grand mean, where the amount of regression toward the grand mean was specified by the within-group slope, $r = .48$. Once we get the parameters for the ANCOVA equation, we can generate the individual posttest scores under the null hypothesis ($\beta_1 = 0$) by using Formula (2):

$$Y_{ij1} = 75.4 + 0 + 0.48 \times Y_{ij0} + e_{ij}$$

where e_{ij} is the random error with mean $M = 0$ and $SD = 13.159$. Once the posttest scores are generated, the group means for males and females could be calculated separately. Note that the standard ANCOVA assumes that disparate groups will regress toward one grand mean under the null hypothesis, with the shrinkage of the distance between the group means specified by the stability coefficient within groups. In contrast, Lord's original paradox specified no shrinkage of the distance between the group means. Most articles contrasting the simple gain score and residual gain score approaches assume that the within-group stability coefficient applies to the expected shrinkage of the differences between the group means, which may be an important contributor to the paradox.

Testing Research Question 3: Simulating Violations of ANCOVA Assumptions

In order to test the consequence of violated assumptions, simulated conditions varied among 1) the normality vs. non-normality of the pretest and posttest scores, 2) the homogeneity vs. heterogeneity of the within-group slopes, 3) homogeneity of variances vs. heterogeneity of variances, 4) different vs. same pretest group means, and 5) two homogeneous sub-groups vs. one heterogeneous group.

3.1. Normality vs. non-normality. Non-normality was simulated by approximating a distribution similar to a zero-inflated distribution, in which 75% of males' and females' pretest and posttest weights were changed to 70 pounds. This was designed to approximate the distribution of depression severity scores in the FFCW data, in which a majority of scores were zero across any two adjacent waves. Seventy pounds

was used because it corresponds to the other scores in a manner similar to how zero corresponds to the non-zero depression scores.

3.2. Homogeneity vs. heterogeneous slopes. Slopes were varied in Lord's paradox, but not in the reversed Lord's paradox. For simulating Lord's paradox, the within-group slopes were varied between zero, 0.48, and 0.96 excluding the case where both slopes were the same. This generated five slope combinations: zero for males and 0.48 for females, 0.48 for males and zero for females, zero for males and 0.96 for females, 0.96 for males and zero for females, and 0.48 for both males and females. Because the reversed Lord's paradox requires regression of the pretest means towards the grand mean to be based on the slope, the slope was fixed to be 0.48 when simulated the reverse Lord's paradox so that the data could fit the null hypothesis of ANCOVA.

3.3 Homogeneous vs. heterogeneous variances. The standard deviation for males was fixed to 15 and varied between five and 15 for females and the standard deviation kept the same for pretest and posttest. This generated the following two variations on standard deviations: 1) males' and females' pretest and posttest have the same standard deviation ($SD = 15$), and 2) males and females pretest and posttest have different standard deviations ($SD = 5$ for females and $SD = 15$ for males). Standard deviations for pretests and posttests were set to be the same.

3.4. Variation of pretest difference between groups. In order to examine the consequence of violating the assumption of independence between treatment and covariates (e.g., the pretest), first, the magnitude of the difference between the mean pretest scores between the two groups was varied from the data designed to fit Lord's (1974) paradox and the reversed Lord's paradox. Then, I simulated equal pretest means

by setting the both pretest group means to be 145 pounds, to compare to the condition when the pretest group means are different (the paradox).

Next, in order to test whether the violation of ANCOVA's assumption on the equality of pretest mean between groups would always lead to inconsistent results, I simulated data varying the pretest and posttest difference. To do that, I set males' pretest and posttest mean to remain equal (160) but varied females' pretest and posttest means separately, from 130, adding 5 pounds for each variation up to a maximum of 160. Females' posttest means started at 130 pounds, and then varied females' posttest by adding 5 pounds each time for a total of seven times. Next, females' posttest means were fixed to 135 pounds, and then females' pretest means were varied by 5-pound increments from 130 to 160 pounds. This was repeated until females' posttest means were fixed to 160 pounds. The total simulation generated a total of $7 \times 7 = 49$ conditions.

The final modification retests Lord's paradox and the reversed Lord's paradox by using the group-centered ANCOVA approach. The simulated data from Lord's paradox and reversed Lord's paradox data was recoded by running ANOVA to get the pretest group mean and pretest residuals after centering on the group pretest mean. Then the pretest group means were subtracted from the posttest score to get the residuals of the posttest scores. Each pretest score was replaced by the pretest residual that centered on the pretest group mean and the posttest outcome was replaced by the posttest residual that centered on the pretest group mean to run the simple gain score approach and the ANCOVA approach.

3.5. In order to generate the most basic version of homogeneous subgroups, I selected the simulated Lord's (1967) paradox data and the revised Lord's paradox, and

divided the data into two sub-groups: one having pretest weights equal or greater than 145 pounds, called relative high weight group, and the other having pretest weights less than 145 pounds, called relative low weight group. Within each subgroup, the simple gain score approach and the residual gain score approach were run to test whether I could get consistent results for the two approaches. The same procedure was used to divide the group and redo the analyses in the reversed Lord's paradox data.

Testing Research Question 4

In order to examine whether the combination of two methods used by Haviland and colleagues (2007) would produce consistent and less biased results, a simple version of propensity score matching with one variable (the pretest weight) is used within each subgroup to create gender groups matched on their pretest propensity score in one sample dataset out of the 1000 obtained. This would be equivalent to propensity score matching with one variable. For doing that, I used Stampf's (2014) package "Nonrandom", a package for conducting propensity score analyses in the R program and followed the example to run the propensity-matching test. Analyses include checking the effect of covariates, generating the propensity score, checking the propensity score distribution, matching samples using the propensity score, checking sample balance after matching, and testing whether results from the simple gain score approach and the residual gain score approach consistency or not. I used the pretest score and the treatment (i.e., gender) to generate propensity scores for each sub-group using optimal matching. The caliper (i.e., matched distance based on standard deviation) for matching was set to be 0.5 (Gu & Fraser, 2014) and the match ratio (i.e., the number of control sample cases matched to the treated sample or the number of treated sample matched to the control sample) was set to

be 2:1. Using the Nonrandom package, the results on testing the propensity score distribution (Figure 3) shows that males and females have overlapping propensity scores. The matched and unmatched sample frequency in Figure 4 indicates that males and females sample are successfully matched in weight level. Males' sample size is twice of females sample size in the relative high weight group and males' sample size is half of females sample size in the relative low weight group. The balance check indicates that in the simulated Lord's (1967) paradox data, the standardized difference (the Cohen's d) changed from a large value (1.28 for the relative high weight group and 1.18 for the relative low weight group) in the original data to a small value (0.25 for the relative high weight group and 0.12 for the relative low weight group) in matched data. In the simulated reversed Lord's paradox data, the standardized difference changed from a large value (1.08 for the relative high weight group and 1.18 for the relative low weight group) in the original data to a small value (0.01 for the relative high weight group and 0.10 for the relative low weight group) in matched data. Note that a standardized difference less than 0.2 indicates that the covariate is balanced adequately between the two experimental groups. The one data analysis from the simulated Lord's paradox and the revised Lord's paradox indicated that a simple version of the combination of the mixture modeling and the propensity score approach is practical in the simulation study. Then I used the 1000 simulated results on Lord's paradox and the reversed Lord's paradox to calculate the average results on the combination of the mixture modeling and the propensity score approach.

Results

The simulated results are summarized in Tables 2, 3, 4 and 5. The first simulation study yielded 48 conditions (Table 2), including 24 for unequal pretest group means (160 pounds for males and 130 pounds for females) and 24 conditions for equal pretest group means (145 pounds for males and for females). Within the first 24 conditions (for unequal pretest group means), 20 of them are variations of Lord's paradox, which assumes that simple gain score analyses fit the correct null hypothesis. Variations include two conditions of normality vs. non-normality, three conditions of homogeneous vs. heterogeneous slopes, and two conditions of homogeneous vs. heterogeneous *SD*. The other four variations start by assuming that ANCOVA fits the correct null hypothesis (the reversed Lord's paradox). Variations include the same two variations for normality and two variations for homogeneity of variance, keeping the same fixed homogeneous slope throughout (0.48). In the second 24 conditions, the pretest means between males and females are fixed to be the same and the other parameters varied similar to the first 24 conditions.

The second simulation study yielded another 49 conditions, varying females' pretest and posttest mean scores (seven variations of females' pretest means times seven variations of females' posttest means), fixing all other parameters to be the same between the two gender groups (Table 3). Each condition had $N = 1000$ Monte Carlo simulations. Table 2 and Table 3 are summaries of the mean of each condition's 1000 simulated results.

Then the simulated Lord's (1967) paradox and the reversed Lord's paradox data were used to test whether I could get consistent results using the simple gains score approach and the residual gain score approach, comparing results from the simulated data, the centered data, and the propensity score samples, which were matched only on the pretest score. Results of these model comparisons are in Table 4 and Table 5.

Research Question 1

The first row in Table 2 replicated Lord's paradox (bold font), showing that males gained significantly more weight than females according to the ANCOVA approach, but there were no significant gender differences using the simple gain score method. The results from the data simulated to reverse Lord's paradox, the other row with a bold font in Table 2, indicated no gender difference using the ANCOVA approach, but females gained more weight according to the simple gain score approach. Analyses of simple gain scores produced the correct result in Lord's paradox, where ANCOVA produced the correct result for reversed Lord's paradox. The correct result is defined herein as the null hypothesis that informed each of the simulated data sets. In both cases, the direction of bias relative to each other is that the ANCOVA results were biased in the direction of the pretest group mean differences compared to the simple gain score approach.

Research Question 3

3.1. The results varied by non-normality primarily due to a greatly reduced magnitude of the bias and reduced power when 75% of the pretest weights and the posttest weights were changed to the low extreme score of 70 pounds from an otherwise normal distribution (Table 2). The effect size changed due to the dilution of the mean by

making a 75% of the samples changed so that both the posttest and pretest weight were changed to be 70 pounds and the other 20% retained their simulated weights for both waves. The dilution of the original simulation made the significant results become non-significant from both the two approaches either when the pretest is not the same for 25% of the cases. The residual plot in Figure 4 indicates that the residuals are not normally distributed when the distribution of the pretest and the posttest weight are not normal.

3.2. The paradox (i.e., getting contradictory results) varies little by homogeneity of the slopes between the two groups when the pretest is different (Table 2). However, smaller average slope coefficients (e.g., .96, .48 and .00) increased the effect sizes from the ANCOVA approach but not from the simple gain score approach when the pretest group means are different. On the other hand, heterogeneous slopes do not change the effect size when the pretest group means are equal (bottom half of Table 2). How the effect sizes change depends on the average slope of the two heterogeneous slopes but not the difference between heterogeneous slopes, but only when the pretest means differ.

3.3. The paradox varies little by homogeneity of the variance between the two groups when the pretest is different (Table 2). However, heterogeneous variances changed the effect sizes from the ANCOVA approach dramatically when the slopes differed, but not from the simple gain score approach when the hypothesis for the simple gain score approach is correct. In contrast, heterogeneous variances did not change the effect sizes either from the ANCOVA approach or from the simple gain score approach when the assumption of no pretest differences were not violated.

3.4 The results indicated that Lord's paradox occurs only when the assumption of equal pretest means is violated (Table 2) no matter whether other parameters are varied or not. The results of 49 conditions (Table 3), which test variations of the distance between two pretest group means indicated that the results from the two approaches were completely consistent (in both the direction and the effect size) only when the pretest group means were equal. Once the pretest group means are different, the results from the two approaches may be inconsistent or consistent in directions, but they will always be inconsistent in effect size. The size of the difference in effect sizes varies proportionately by the size of the differences in pre-test group means, $b_1 - d = (1 - slope)(\bar{Y}_{20} - \bar{Y}_{10})$. Where, d is the effect size of the treatment effect using the simple gain score approach and b is the effect size of treatment using the residual gain score approach.

Table 4 presents the results from the group-centered ANCOVA approach, which centered the pretest and posttest outcome scores around the pretest group means. The two approaches are compared for Lord's paradox data and reversed Lord's paradox data. By centering the ANCOVA on the pretest group means, the results from the simple gain score approach and the residual gain score approach using the centered data are consistent with each other, but duplication the simple gain score approach using the original data. From simulated Lord's paradox data, the gender effects are $d = -0.01$ (not significant) using the simple gain score approach, $b = -15.60$ ($p < .001$) using ANCOVA, and $b = -0.01$ using the group-centered ANCOVA approach. From reversed Lord's paradox simulated data, the gender effects are $d = 15.61$ ($p < .001$) using the simple gain score approach, $b = 0.02$ (n.s.) using the ANCOVA, and $b = 15.61$ ($p < .001$) using the group-centered ANCOVA approach. Thus, the group-centered

ANCOVA eliminates the contradictory results from Lord's paradox and reversed Lord's paradox, but it produced the correct causal effect only for Lord's paradox, that is with data designed to fit the null hypothesis of no treatment effect according to simple gain scores. The results on the reversed paradox indicated that although the effect size from the simple gain score ($d = 15.61$) is the same as the effect size from the group-centered ANCOVA approach ($b = 15.61$), the group-centered ANCOVA ($t(d) = 18.76$) has more power than the simple gain score approach ($t(b) = 16.17$).

3.5. In order to test a simple version of propensity score by matching samples on the pretest weight, the data for both the simulated Lord's paradox and the reversed Lord's paradox were split into two parts. One part has the pretest weight greater and equal to 145 pounds, called the high weight group, and the other part has the pretest weight less than 145 pounds, called the low weight group. The top part of Table 5 shows the results from only one of the 1000 obtained datasets, and the bottom part of Table 5 shows the average results from all 1000 datasets. Results from the split simulated original paradox data indicated that the pretest means for males and females are still significantly different for both the high weight sub-group ($d = -13.02$ from one dataset and $d = -11.45$ from all 1000 datasets, $ps < 0.001$) and the low weight sub-group (-11.21 from one dataset and $d = -11.45$ from all 1000 datasets, $ps < 0.001$). The results from the two approaches are consistent on the direction of the effect, but the effect sizes are not close ($d = -7.26$ using simple gain scores and $b_I = -13.67$ using residual gain scores from one dataset, and $d = -9.60$ using simple gain score and $b_I = -15.55$ using residual gain score from all 1000 dataset, all $ps < 0.0001$) in the high weight sub-group. In addition, the results from the two approaches are consistent in the low weight sub-group on the effect direction but the

effect sizes are not close ($d = -12.37$ using simple gain scores and $b_I = -18.75$ using residual gain scores from one dataset, and $d = -9.72$ using simple gain scores and $b_I = -15.64$ using residual gain scores from all 1000 datasets, all $ps < 0.0001$).

Results from the split simulated reversed paradox data indicated that the pretest means for males and females are still significantly different for both the high weight sub-group ($d = -11.60$ from one dataset and $d = -11.44$ from all 1000 datasets, $ps < 0.001$) and the low weight sub-group (-10.29 from one dataset and $d = -11.46$ from all 1000 datasets, $ps < 0.001$). The results from the two approaches in the high weight sub-group are inconsistent on the effect sizes ($d = 7.37$, $p < 0.001$, using simple gain scores and $b_I = 0.72$, $p > 0.10$, using residual gain scores from one dataset, and $d = 6.00$, $p < 0.05$, using simple gain scores and $b_I = 0.07$, $p > 0.10$, using residual gain scores from all 1000 datasets). The results from the two approaches in the low weight group are inconsistent on the effect sizes ($d = 3.41$, $p < 0.10$, using simple gain scores and $b_I = -1.05$, $p > 0.10$, using residual gain scores from one dataset, and $d = 6.00$, $p < 0.05$, using simple gain scores and $b_I = 0.01$, $p > 0.10$, using residual gain scores from all 1000 datasets).

Research Question 4

Research question #4 is to test whether I can get consistent results from the simple gain scores and the residual gain scores by a simple version of propensity score method, matching samples on the pretest weight. Results from the matched split simulated data indicated that the pretest means for males and females differed only marginally for the high weight sub-group using one dataset ($d = -1.51$, $p = 0.08$) in the original Lord's paradox and did not differ significantly in all others datasets. Using the matched sample,

the results from simple gain score approach and the residual gain score approach are consistent on direction and close on effect size either using one dataset or using the average score of 1000 simulated datasets. Results from the simulated Lord's paradox data showed that in the high weight sub-group, the effect sizes are $d = -11.77$ from one dataset and $d = -15.00$ from all 1000 datasets using simple gain scores and $b_I = -12.74$ from one simulated dataset and $b_I = -15.55$ from all 1000 simulated datasets, using residual gain scores, all $ps = < 0.001$. In the low weight group, the effect sizes are $d = -18.75$ from one dataset and $d = -15.10$ from all 1000 datasets using simple gain scores and $b_I = -19.11$ from one dataset and $b_I = -15.63$ from all 1000 datasets using residual gain scores, all $ps = < 0.001$. Results from the simulated reversed Lord's paradox data showed no gender differences, consistently from the simple gain score approach and the residual gain score approach either using the one simulated dataset or using the average score of the 1000 simulated datasets.

Discussion

The purpose of the simulations is to reveal inconsistent and potentially biased results from the simple gain score approach and the residual gain score approach in nonrandomized studies, illustrated by Lord's paradox, to explore possible reasons for the biased and inconsistent results, and to introduce possible corrections for any biasing factors identified. Because differences in pretest group means were most closely associated with inconsistent results, I used the modified group-centered ANCOVA to remove mean group differences in the pretest scores. I first demonstrated that pretest group differences would lead to inconsistent results between the simple gain score approach and the residual gain score approach. This confirmed that at least one statistical

approach is biased when pre-test differences existed between the treatment group and the comparison group in nonrandomized studies.

I then simulated a series of datasets to test violations of simple gain score's assumptions and/or ANCOVA's assumptions: normality, homogeneous variance, homogeneous slope, and independence of treatment and covariate/pretest. I found that violations of the assumption of homogenous variance do not apparently contribute to the paradox but the size and heterogeneity of slopes influence the effect size of the ANCOVA approach when the pretest means differ. The ANCOVA effect sizes get larger as the average slope coefficient gets smaller from 1.00 to .00, although I only compared .48 and .24. ANCOVA is known to produce the same effect size as the simple gain score approach when the slope coefficient is 1.00. The more the average slope coefficient gets smaller than 1.00, the greater the discrepancy between the effect sizes from ANCOVA and the simple gain score approach. When the average slope coefficient is held constant (e.g., at .48 in Table 2), slope heterogeneity does not change the effect size by itself. However, the combination of heterogeneity of both slopes and variances can change the effect sizes a lot. Heterogeneity of variance influences ANCOVA effect sizes only in combination with slope heterogeneity. The result suggested that the test of homogeneity of slope and variance should be carried out whenever the ANCOVA approach is applied. The results also support previous findings that when covariate means are different between the comparison group and the active treatment group, slope heterogeneity will increase the type I error rate (Hollingsworth, 1980; Huitema, 2011). At the same time, the violation of the assumptions of normality that were tested distort the results when most of the data does not change due to staying at a floor level. This result

may apply to analyses of depression severity in the next chapter, because most scores are unchanged from the minimum possible score. On the other hand, changing most of the weights to a low value of 70 resulted in consistent results from both statistical approaches that females gained more weight than males in the simulated data for reversed Lord's paradox. Violation of the assumption of independence between treatment and covariate results in inconsistent results from the two approaches. When the simulations had no group difference in the pretest/covariate means, the effect sizes were consistent no matter how other parameters varied.

Next, I varied the pretest mean differences between the comparison group and the active treatment group to examine how much difference would lead to the inconsistent results, through simulations. The results indicated that even slight differences between the comparison group mean and the active treatment group mean on the pretest will lead to inconsistent effect sizes between the two approaches, although their direction may be consistent sometimes. The difference of effect sizes between the two approaches is $b_1 - d = (1 - slope)(\bar{Y}_{20} - \bar{Y}_{10})$, where d is the size of the treatment effect using the residual gain score approach and b_1 is the effect size of treatment using the residual gain score approach. This equation implies that the two approaches will produce consistent results ($b_1 = d$) only if the slope used to predict shrinkage of the distance between the group means is equal to one, making $1 - slope = 0$, or if there is no pretest difference between the two groups, $\bar{Y}_{20} - \bar{Y}_{10} = 0$. This assumes equal variances in the pretest and posttest distributions. Table 3 also includes variations in which the two approaches produce contradictory signs of their effect sizes, sometimes significantly in opposite directions.

Using group-centered ANCOVA to recode the data to remove the group difference in pretest means did produce consistent results from the simple gain score approach and the residual gain score approach. This suggests that group-centered ANCOVA has one important advantage over ANCOVA, in that it satisfies the assumption of no treatment difference on the covariate. It retains the other purpose of ANCOVA, which is to reduce the residual variance to be explained by controlling statistically for extraneous covariates that predict the outcome (Huitema, 2011). But the consistent effect sizes from group-centered ANCOVA are nearly identical to the original effect sizes from the simple gain score approach either using the simulated Lord's (1967) paradox data or using the simulated reversed Lord's paradox. This is correct for the simulation of Lord's paradox since the data fit the null hypothesis of simple gain score but very biased for the simulation of reversed Lord's paradox since the data fit the null hypothesis of ANCOVA.

Using the matched sample, the effect size from the simple gain score and the residual gain score are almost the same using the average score of the 1000 simulated data and very close to each other from one simulated data, especially when the pretest differences became non-significant between males and females in the relative low weight group. Matching cases on the pretest score produces consistent results, just like group-centered ANCOVA. But the consistent effect sizes from the matched samples are nearly identical to the original effect sizes from the residual gain score approach. In contrast to the results of group-centered ANCOVA, the results from the matched sample is very biased for the simulation of Lord's paradox, but is unbiased for the simulation of reversed Lord's paradox.

So I have shown two methods for making the pretest group means equal, both of which produce consistent results for the simple gain score approach and the residual gain score approach. The problem is that these two sets of consistent results differ from each other. In fact, they are just as far apart as the original contradictory results from the two approaches. Therefore, neither set of consistent results can be counted on to produce less biased causal estimates than the original contradictory results. Either result could be unbiased, but only if it corresponds to the approach with the correct unbiased null hypothesis. It may be that additional covariates can clarify which approach is less biased or reduce the bias in one or both of them.

Overall, the results suggest that satisfying the assumption of ANCOVA that the covariate be independent of the treatment is the key for ensuring consistency of the two approaches. Furthermore, it can be pointed out that although the simulation results indicated that when the pretest is the same the results from the two approaches are consistent, it is no guarantee that the results will be unbiased in nonrandomized studies. Analyses of reversed Lord's paradox showed that group-centered ANCOVA produced consistent results that are just as biased as the most biased analysis of those data. Analyses of Lord's original paradox showed that matched samples produced consistent results that are just as biased as the most biased analysis of those data. Another possible reason is that beside the pretest, other covariate differences between the comparison group and the active treatment group on the baseline also could violate the assumption of ANCOVA and lead to biased results (Van Breukelen, 2006). In addition, although we could simulate data to fit the null hypothesis of differences-in-differences or the null hypothesis of ANCOVA and analyze which approach is more biased, when analyzing

real data, there is often no way to know which approach is less biased. As such, neither the group-centered ANCOVA nor pretest-matched samples are able to reveal the true treatment effect.

Why do both sets of consistent results remain potentially biased? Consider group-centered ANCOVA first. By only centering both pre-test and post-test scores on the within-group pretest means, we are assuming that the group means will retain the same difference between them at the post-test as at the pre-test, according to its null hypothesis of no treatment effect. This matches the standard that is implicit in predicting simple change. However, the pretest outcome is not equal between the comparison group and the active treatment group in reality and we do not know how much the difference could actually affect outcomes apart from the treatment. Thus, although the pretest/covariate difference between the comparison group and active treatment group is statistically removed and the consistency of the estimation of the treatment effect is guaranteed, the results could remain biased in nonrandomized studies.

What about matched samples? The purpose of matching is to equate the samples on the matching variables, in this case the pretest scores, which is identical to the purpose of controlling for a covariate, in this case to equate the groups on the pretest.

Advanced models explored whether a combination of the mixture modeling and propensity score methods could achieve consistent results that are successful in reducing bias, especially with matching on more covariates than just the pretest. Since the data have been separated into two subsample datasets, there was not enough evidence to make suggestions on whether we could get less biased results from a combination of mixture

modeling and propensity score methods, when based on more covariates. The effectiveness of the combination of the mixture modeling and the propensity score method was further examined in the real data set in the next Chapter.

CHAPTER IV

STUDY II: THE FRAGILE FAMILY AND CHILD WELL-BEING DATA

The purpose of using the FFCW data is to apply lessons from the simulated results based on Lord's paradox and to examine whether improved models could produce consistent and less biased results. First, I investigated whether Lord's (1967) paradox applies to treatments for depression, starting with a comparison of the simplest applications of the two traditional methods across two waves, the simple gain score approach and the residual gain score approach. These simple two-wave analyses investigated whether the two approaches lead to the contradictory results. Next, I compared the results of the cross-lagged panel model and a linear growth model across three waves, which incorporates the two traditional adjustment methods, respectively. Then, I diagnosed the problems by testing the assumptions of the simple gain score approach and ANCOVA to examine whether the data meet assumptions such as normality, homogeneity of slopes, and one homogeneous group with the same developmental trend in depression severity, which was used to compare the active treatment group and the comparison group. Since the simulation results indicated that

removing the pretest difference by group-mean centering could achieve consistent results that retain the bias of the simple gain score approach, I assume the analyses from the FFCW data will get the same conclusion. We cannot be certain about achieving an unbiased causal estimate in real data, but we assume that significant adverse average effects of established treatments for depression are due to a bias in the causal estimate. In the next several steps, I will use model comparisons to examine whether I could further get less biased results. Fourth, based on the diagnostics about assumptions for analyses of simple gain scores and ANCOVA assumption diagnoses from the previous step and the simulation, model improvements will be conducted step by step within each homogeneous subgroup, after employing mixture modeling to identify homogeneous subgroups to determine whether the result from the two approaches could get closer to each other. Finally, propensity score adjustments will be introduced to determine whether I can get more consistent and less biased results from matched experimental groups. Haviland and colleagues (2007) obtained consistent results across both types of baseline adjustment approaches when they combined a version of mixture modeling with propensity-score methods. I will follow their example of a combination of the two approaches to determine whether I can get more consistent and less biased results across those two baseline adjustment methods from their combination of the two approaches, consistency that would be found in idealized randomized studies.

Methods

Participants

The real data for analyses in the present study are from the FFCW data, which started with baseline data for mostly unmarried couples with children born from 1998 to 2000 in 20 large cities of the United States. The FFCW study collected information including household characteristics, physical and mental health, and parenting behaviors through in-person interviews when the children were born, and follow-up telephone interviews when the children were approximately 1, 3, 5, and 9 years old. Starting when the children were age 3, in-home interviews and observations were conducted, collecting information across multiple domains of parenting, the home environment, mother-child interactions, and the child's cognitive and emotional/behavioral development. For this study, I drew demographic variables (e.g., gender, poverty ratio, and maternal education), two types of treatment for depression (medication treatment and psychotherapy treatment), and the four waves of maternal depression symptoms from the core telephone interview, and potential confounding variables from both telephone interviews and in-home interviews.

Measures

Treatment. Two types of treatments were operationalized treatment for depression: medication treatment and psychotherapy treatment. Measures for medication treatment come from two questions at each wave from Wave 2 to Wave 5. For medication treatment, for example, mothers were asked “during the past 12 months, did you receive counseling or therapy for personal problems, for example, feelings of

depression, worry, alcohol, or drug use problems?” If the answer was yes, they were further asked what conditions they took medication for. Options included diabetes, asthma, high blood pressure, depression, anxiety, attention deficit, pain, seizures or epilepsy, and others. If mothers answered “yes” to the first question and indicated they took medication for depression, the score for the medication treatment was coded “1,” and others cases were coded “0.” The measure for psychotherapy treatment for depression comes for another two questions. Mother’s reported on whether they received counseling/therapy for personal problems in the past year. If the answer was “yes”, they were asked whether the counseling/therapy was for depression, anxiety, attention problems, alcohol problems, drug use problems, or something else. Mothers’ responses indicating that they received psychotherapy for depression were coded “1,” and other cases were coded 0.

Depression. Maternal depression symptoms were assessed by maternal self-reports about symptoms of a Major Depressive Episode (MDE), derived from the composite International Diagnostic Interview- Short Form (CIDI-SF), Section A (Kessler, Andrews, Mroczek, Ustun, & Wittchen, 1998). The CIDI is a standardized instrument for assessing mental disorders. Participants responded to several questions step by step (Figure 6). First, they indicated in two stem questions whether in the past year they had feeling of being sad, blue, or depressed that lasted for two weeks or more (Question J5) or lost interest in most things (Question J9). If yes to either question, they said whether the symptoms lasted all day long, most of the day, about half of the day, or less than half the day (J6 and J10). If the answer was about half of the day or more, they said whether the feeling occurred every day, almost every day, or less often for a two-

week period (J7 and J11). If the answer was at least almost every day for two weeks they answered another set of seven items on whether they were: losing interest (J8), feeling tired (J12), changing in weight (J13 and J13a), having trouble sleeping (J14 and J14a), having trouble concentrating (J15), feeling worthless (J16), and/or thinking about death (J17). Those questions are used to assess whether participants meet the depression criteria for a DSM-IV depressive episode. To meet the diagnosis requirement for Major Depression (MD), the participant should endorse either one of the two stem questions (i.e., feel sad, blue, or depressed for Question J5, or lose interest in most things for Question J9) lasting at least half of the day and occurring at least almost every day during a two-week period. Once they meet the criteria on one of the two stem questions, the seven follow-up questions and the first question (J5) are used to calculate the MD score (range: 0-8), which counts symptoms for determining the likelihood of a diagnosis of Major Depression. Positive responses to questions “lose interest,” “feeling tired,” “having trouble in concentration,” “feeling worthless,” “thinking about death” and the first question (J5) were coded as “1” for each question. Weight changes equal to or more than 10 pounds were coded as “1” and having trouble sleeping at least nearly every night was coded as “1.” The sum of the eight questions is computed to yield the MD score. An MD score equal to or greater than “1” indicates the participant has one or more depression symptoms. In this study, more than 75% of mothers’ MD scores are zero. In order to distinguish those mothers having some depression symptoms that occur less than half of the day or less than almost every day during a two-week period from those mothers not having any depression symptoms, a 13-point scale was created to measure mothers’ depression severity. If a mother’s MD score was above zero, her depression severity

score became her MD score plus 4. If a mother's MD score was zero, I added the positive responses to question J5, J6, J9, and J10 to yield the other 4 points in the 13-point scale. Mothers answer "yes" to question J5 that they have felt sad and blue, answer "yes" to question J9 that they had lost interest, and respond to question J6 and to question J10 that the symptom occurred more than half of the day are coded "1" for each question. This was necessary to discriminate between mothers who indicated no depressive symptoms at all from those who indicated only sub-threshold symptoms for Major Depressive Disorder. The sum of the four dummy codes yielded the depression severity score for those mothers whose MD score are zero. Table 6 and Figure 7 show the depression severity scores and their frequency for Wave 3 and Wave 4 depression symptoms.

Covariates. Propensity-score adjustments are only as good as the covariates that contribute to the propensity scores. As mentioned in the literature review, either the potential confounders or the true confounders should be prioritized for covariate selection. In the case of depression, published studies include much more data on predictors of depression than of treatment for depression, such as mothers' age, inter-parental conflict, social economic status, previous trauma experience, and stress of being a parent. Previous studies indicated that depression results from an interaction of biological, psychological, and sociological factors (Beck, 1996; Depue, 1979). Factors include but are not limited to status attributes (age, sex, education, marital status, income, and race), personal resources (psychological state, coping, social support, social connections, low self-esteem, parental stress, child temperament problems), health conditions (smoking, using alcohol, and using drugs), and previous experience (trauma, stress events, family history of depression) (Kaplan, Roberts, Camacho, & Coyne, 1987).

Considering these factors and the previous studies on depression using the FFCW data, I selected 27 variables (age, sex, education, marital status, poverty ratio, financial harshness, financial harshness for medication, race, foreign born status of mothers, smoking, alcohol use, and drug use, number of children, intimate partner domestic violence or aggression, parenting stress, mother's health condition, child's health condition, child aggression, previous depression symptoms) that cover the area of biological, psychological, and sociological area as potential confounding covariates. Each potential confounding covariate is used to test its correlations with depression and with treatments for depression using the FFCW data. The results in Table 7 list both types of correlations and the tests of whether a covariate is a true confounder (Austin, 2011) and should be included in the covariate list for propensity score analysis. The point is, to be a true confounding variable, a covariate must be related to both treatment assignment and the outcome. In order to increase the probability of matching, the significant level for correlation tests will be set to be 0.05. The final 16 covariates that were included in the propensity score analysis are: 3 previous wave of depression scores, smoking, financial hardship, domestic violence, child externalizing problems, parental stress, drug used, mother foreign born, child health conditions, social support, intimate partner support, cohabitation status, and mothers' health condition (quadratic effect).

Results

In this chapter, using the FFCW data, I first verify the contradictory results from the simple gain score approach and the residual gain score approach, first using two-wave data. This replicates the simulation results based on Lord's paradox. Next, I show contradictory results from those two approaches using three-wave data, comparing the

cross-legged panel model and the linear latent growth model. I then test whether the assumption of ANCOVA are violated in the FFCW data, including normality of the residual outcome scores, homogeneous slopes, and independence of treatment condition and pretest scores. If an assumption is violated in a test, that assumption will be tested again in the next step to investigate whether the assumption violation is overcome or not when the model is improved. Next, I test whether the longitudinal depression data set are composed of more than one heterogeneous trajectory subgroup and how many trajectory subgroups there are. Once the trajectory subgroups are identified, I test whether the assumption violations found in the previous step are overcome within each homogeneous trajectory subgroup. Finally, I use propensity score methods to match samples in the treatment group and the control group within each trajectory subgroup. Using matched samples within each trajectory subgroup, I test whether the assumption violations found in the previous step are overcome in this step. The main point is to determine the steps at which the model improves in satisfying the above-mentioned assumptions, in getting consistent results from the simple gain score approach and the residual gain score approach, and ideally in demonstrably less biased results, by using the matched sample subset data within each trajectory subgroup. At the same time, the results using the matched sample within each trajectory subgroup are compared to the results from the group-centered ANCOVA approach using the total sample within each trajectory subgroup.

Software programs for the present study are Mplus Version 8.0, Stata Version 15, and R Version 3.5.0. Multiple imputation (Syntax see Appendix C) is used to handle missing data, because all covariates had a certain level of missing data (e.g., the mothers'

depression on Wave 5 had 28% of missing, intimate partner support had 25.72% of missing), after identifying the trajectory subgroups. In order to increase power, all 27 potential covariates were included in the imputation. It was assumed that, when using multiple imputation estimation, missing are at random or random after including other variables in the analysis. Density plots in Figure 8 show the patterns of the non-imputed data (blue) and the ten imputed data sets (red), indicating that the original data is well represented by the imputed data. The average scores from the ten imputed datasets are used in propensity score matching. Analyses results using FFCW dataset are in Table 8 to 11 and Figure 11 to 20.

Research Question 1

Research Question #1 asked whether it is possible to get consistent results from the simple gain score approach and the residual gain score approach in two-wave analyses without other covariates. The results in Table 8 indicate that both treatments for depression reduced depression symptoms according to the simple gain score approach, $d = -2.31$ for psychotherapy and $d = -1.87$ for medication, but increased depression symptoms according to ANCOVA, $b = 1.74$ for psychotherapy and $b = 1.79$ for medication, all $ps < .001$.

Consistent with the simulation results, group-mean-centered ANCOVA produced consistent results. When the pretest and posttest depression scores were both centered around the pretest depression group means, the results from group-centered ANCOVA ($b = d = -2.31$ on psychotherapy and $b = d = -1.95$ on medications with $ps < .001$) were consistent with each other and nearly identical to the results using the simple gain score

approach above. Consistent with the simulated analyses, group-centered ANCOVA eliminated the contradiction between the two approaches, but its results are unbiased only if the original gain score approach was unbiased.

Research Question 2

To test whether contradictory results generalized to analyses across three or more waves, Research Question #2 tested whether I could get consistent results from the cross-lagged panel model and the latent growth model from a three-wave data analysis without any covariates. The structural equation models for cross-lagged panel analyses and two-step latent growth models across Waves 3, 4, and 5 are shown in Figures 9 and 10.

Although standard latent growth models typically predict one linear slope from the first to the last wave, the two-slope latent growth model in Figure 10 is designed to be more similar to a cross-lagged panel by predicting simple change scores between adjacent waves. The intercept is modeled as usual (all loadings set to 1), but Slope 1 specifies simple change from Wave 3 to Wave 4 (with loadings set at -1 and 0), whereas Slope 2 specifies simple change from Wave 4 to Wave 5 (loadings set to 0 and 1). The model then estimates the effect of treatment for depression at one wave (Wave 3 or 4) on the simple change in depression severity from that wave to the next wave.

According to the cross-lagged panel models shown in Figure 9, treatment at any wave looks significantly harmful by increasing depression severity at the next wave, controlling for the preceding depression severity score: $b = 0.70$ (Wave 3 psychological treatment predicting Wave 4 depression), $b = 1.60$ (Wave 4 psychological treatment predicting Wave 5 depression), $b = 1.22$ (Wave 3 medication treatment predicting Wave

4 depression), and $b = 1.40$ (Wave 4 medication treatment predicting Wave 5 depression), all $ps < .05$. In contrast, using the 2-slope latent growth models in Figure 10, treatment at any wave looks helpful in reducing depression severity from that wave to the next wave: $b = -2.98$ (Wave 3 psychological treatment predicting decreasing depression from Wave 3 to Wave 4), $b = -0.62$ (Wave 4 psychological treatment predicting decreasing depression from Wave 4 to Wave 5), $b = -2.15$ (Wave 3 medication treatment predicting decreasing depression from Wave 3 to Wave 4), and $b = -1.05$ (Wave 4 medication treatment predicting decreasing depression from Wave 4 to Wave 5), $ps < .05$. Thus, the same contradictory results found in 2-wave analyses generalize to 3-wave analyses.

Research Question 3

Research Question #3 asks whether the contradictory results and corresponding biases are due to violations of assumptions, especially in ANCOVA. Evaluated assumptions include normally distributed outcome scores and residuals, equivalent within-group slopes from the depression severity (pre-test) at Wave 4 to the depression severity outcome (post-test) at Wave 5, equivalent group mean scores on the covariate (pre-test at Wave 4), and a homogeneous group rather than a mixture of heterogeneous groups. The assumptions were tested one by one separately and then in combination.

Test 3.1. The first diagnostic test is whether the distribution of depression severity is normal. The depression severity frequencies at Wave 2, Wave 3, Wave 4, and Wave 5 in Figure 7 indicated that depression severity is not normally distributed but has an inflated number of zeroes with means and standard deviations of $M = 1.85$ and $SD = 3.68$ for Wave 4, and $M = 1.84$ and $SD = 3.68$ for Wave 5. Except for the excessive numbers

of zeroes, the distribution of depression severity scores looks somewhat continuous and is not skewed at either extreme. Therefore the analyses used a modification of a hurdle model. A hurdle model handles the excessive zeroes by separating the analysis into two parts. The first part of a hurdle model uses logistic regression to predict zeroes vs. non-zeroes at Wave 5. The second part predicts the non-zero scores by themselves because they got over the first “hurdle.” I modified the hurdle model by defining the hurdle as those with consistent zero scores at both Wave 4 and Wave 5. This included zero scores at Wave 5 in the continuous part of the hurdle model for those who had non-zero scores at Wave 4. Otherwise a hurdle model would exclude those who improved the most from the continuous part of the analysis (the second part). Cases with zeroes at Wave 4 were also retained if they had non-zero scores at Wave 5, so that the distribution of depression severity would be similar at Waves 4 and 5. Thus the modified hurdle model used logistic regression to predict cases that had consistent zero scores at Waves 4 and 5 vs. those that had non-zero depression severity scores for at least one of those two waves. The remaining cases were then analyzed in the continuous part of the modified hurdle model. Analyses were expected to be robust for non-normality in the continuous part of the modified hurdle model, based on the Central Limit Theorem.

The logistic model for predicting depression symptoms vs. consistently zero depression symptoms showed that those mothers who were in psychological treatment at Wave 4 had odds of having depression symptoms that was 25.88 times higher than those mothers who were not in psychological treatment. Also, for those mothers who were in medication treatment at Wave 4, their odds of having depression symptom was 9.03 times higher than those mothers who were not in medication treatment, $p < 0.001$. For the

non-zero part of the analysis, results from the simple gain score approach and the residual gain score approach were compared. Using the residual gain score approach, both psychological treatment ($b = 1.63, p < 0.001$) and medication treatment ($b = 1.98, p < 0.001$) appeared to be harmful for depression, whereas using the simple gain score approach, both psychological treatment ($d = -2.70, p < 0.001$) and medication treatment ($d = -2.27, p < 0.001$) appeared to be beneficial for depression. Thus, the contradictory results remained when the violation of the normality assumption was reduced.

Test 3.2. I tested the second assumption concerning whether the active treatment group and the comparison group have the same slope by testing the Treatment \times Pretest interaction effect, excluding on posttest depression severity, excluding the mothers with consistent zero scores across both waves. The results showed that the interaction effect was significant for the psychological treatment ($b = -0.12, p = 0.03$), whereas the interaction effect was not significant for the medication treatment ($b = -0.02, p = 0.77$). This indicated that the slope for psychological treatment are heterogeneous for the treatment group and the control group, whereas the slope for the medication treatment are homogeneous for the treatment group and the control group. The Johnson-Neyman's approach (D'Alonzo, 2014; Huitema, 2011) is recommended when the slopes are heterogeneous, but considering interactions are beyond the scope of this dissertation and was not tested herein.

Test 3.3. The third diagnosis is to test whether the distribution of the residuals for both the psychological treatment and the medication treatment are normal or not. The histogram plot in Figure 11 indicated the residuals for both the psychological treatment and the medication treatment approximated a normal distribution better than the original

distribution of depression severity (Figure 7d). Nonetheless, the distribution of residuals still includes an excess frequency of minimum scores, although that frequency now equals the peak of the normal part of the distribution, whereas it was 16 times as frequent as that peak in Figure 7d. Additional steps to make the distribution more normal seem problematic, because it would systematically remove those who improved from non-zero to zero symptoms from Wave 4 to Wave 5.

Test 3.4. The aim of the fourth diagnostic is to test whether the pretest mean scores are the same across the treatment and control subgroups. An independent samples *t*-test is used for this analysis. As expected, the pretest mean scores were significantly different between the active treatment group and comparison group. For the psychological treatment, the pretest mean difference between the active treatment group and the comparison treatment group was 5.79 and for the medication treatment, the pretest mean difference between the active treatment group and the comparison treatment group was 5.42, $ps < 0.001$. At a minimum, some later steps are likely to reduce the discrepancies between pretest means, which could produce more consistent results across the two approaches for analyzing change scores.

Test 3.5. The fifth diagnosis tested whether the FFCW data represents one homogeneous group on developmental trends in depression symptoms before getting treatment. If not, how many sub-groups are there? Depression severity in Wave 2 to Wave 4 data were used to identify sub-groups using mixture modeling running in M-plus Version 8.0. Mixture modeling is used to test the possibility of subpopulations within one overall population. In the present study, linear latent class analyses, a specific type of mixture modeling for longitudinal data, were used to run **2-sub-group, 3-sub-group, and**

4-sub-group models separately. Information Criteria such as Akaike (AIC), Bayesian (BIC), and Sample-Size Adjusted BIC (SSA-BIC), and model comparisons using the Vuong-Lo-Mendell-Rubin Likelihood Ratio Test (LMR) results are employed to determine which model is superior. If the first three criteria provide opposite suggestions, the BIC result will be considered primary (Nylund, Asparouhov, & Muthén, 2007). If the model comparisons based on log-likelihood test results indicate no significant differences between two adjacent models, the solution with fewer groups will be considered better. If the model comparisons of log-likelihood test results indicate significant difference between two adjacent models, the model with more groups is better, and further tests adding one more group will be conducted. Moreover, if a solution includes one subgroup with less than 5% of the overall sample, the solution is not considered.

The distribution of depression severity scores in the four waves of data indicated a zero-inflated distribution, in that about 70% of mothers did not have any depression symptoms for each wave (Figure 7). For those mothers with non-zero depression scores, the distribution approximates normality according to the Center Limit Theorem. In order to fit the distribution of the data, Latent Linear Growth using Negative Binomial Hurdle Model was used to estimate trajectory groups across Wave 2 to Wave 4 (Syntax see Appendix D). Three models were fit to the data, specifying two latent classes, three latent classes, and four latent classes. For each model, replication of the best log-likelihood was verified with varying start values to avoid local maxima. In order to avoid singularity of the information matrix, one or two of the main parameters (intercepts, slopes for the logistic part and for the linear regression part) were fixed to be zero in one or two subgroup. Furthermore, the variance of the intercepts and slopes are fixed to zero since it

assumes that, within each subgroup, mothers have similar depression patterns with little variations (Haviland et. al., 2007).

Results are summarized in Table 9 and Table 10. The top half of Table 10 shows the Information Criteria, entropy, likelihood ratio tests, and the test results of each model and the bottom half of Table 10 shows the sample size of each subgroup and the average latent class probabilities for the most likely latent class membership for each model. The information criteria (AIC, BIC, and SSA-BIC) indicated that the three-group model was better than the two-group model and that the four-group model was better than the three-group model. Also, the LMR test indicated significant difference between the three-group model and the two-group model and significant difference between the three-group model and the four-group model. All the above four criteria indicated the four-group model is better than the three group model and the three group model is better than the two group model. However, the four-group model has one subgroup with only 1.09% of the sample, indicating the sample size for that subgroup is too small for analysis. Thus, the three-group model was used in the further analyses.

The results in Table 10 from the three trajectory subgroups show that 65.88% of the mothers are in the low depression group, 26.30% in the high depression group with a peak depression score at Wave 3, and 5.81% of mothers in the medium depression group with a peak depression score at Wave 2. Within the low depression subgroup, 1.88% of mothers received psychological treatment and 1.42% of mothers received medication treatment at Wave 4. The mean depression scores from Wave 2 to Wave 4 for the low depression subgroup are 0.22, 0.30, and 0.36. Within the High depression subgroup, 12.23% of mothers received psychological treatment and 10.00% of mothers received

medication treatment at Wave 4. The mean depression scores from Wave 2 to Wave 4 for the High depression subgroup are 4.48, 5.74, and 4.84. Within the Medium depression subgroup, 9.74% of mothers received psychological treatment and 7.87% of mothers received medication treatment at Wave 4. The mean depression scores from Wave 2 to Wave 4 for the Medium depression subgroup are 4.41, 3.61, and 3.64.

Test 3.6. Once the trajectory subgroups was identified, it was expected that the depression patterns within each trajectory subgroup would be more similar and some of the violated assumptions may not be the case anymore. Thus, the next step reconsiders the assumption violations within each of the three subgroups. The assumptions to be reconsidered are the assumption of normally distributed outcomes and residuals, homogeneous slopes, and homogeneous pretest group means. Additional tests of those assumptions were conducted within each subgroup to examine whether the evidence of apparent adverse effects of treatments for depression disappears, and whether I can get consistent results from the two baseline adjustment methods. I used the imputed data that were generated from the multiple imputations to replace missing data.

Several steps were used to test these assumptions in the trajectory subgroups. First, I tested whether the outcome within each trajectory subgroup are normality distributed. Figure 12 indicated that although the low depression subgroup contains the majority of the zero depression scores, the distribution of each of the three trajectory subgroups still shows a zero-inflated distribution. Second, I tested whether the residuals within each trajectory subgroup is normality distributed. Figure 13 indicated that normally distributed residuals are best approximated in the High trajectory subgroup. Third, within each subgroup, I tested whether the pretest depression mean is different for

the active treatment group and the comparison group. The results (Table 11) indicated that all the pretest means differ between the active treatment and the comparison group: in the low depression trajectory subgroup $d = 2.74$ ($p < 0.001$) for psychological treatment and $d = 2.58$ ($p < 0.001$) for medication treatment, in the Medium trajectory subgroup $d = 4.63$ ($p < 0.001$) for psychological treatment and $d = 4.30$ ($p < 0.001$) for medication treatment, and in the High trajectory subgroup treatment $d = 3.17$ ($p = 0.002$) for psychological treatment and $d = 1.94$ ($p = 0.06$) for medication treatment. These pretest group means are closer than they were in the full sample: $d = 5.79$ for psychological treatment and $d = 5.42$ for medication treatment. Forth, within each subgroup, I tested whether the slopes of the active treatment group and the comparison group differs from each other. The results indicated no significant Treatment \times Pretest interaction effect within any trajectory subgroup, indicating the slopes are homogeneous. Overall, the assumption tests show that, at this point, the assumption of normality and the assumption of independence of treatment condition and pretest scores are still violated within each trajectory subgroup. These two assumption violations may be resolved when testing Research Question # 4. After testing these assumptions test, I tested whether I could get closer results from the simple gain score approach and the residual gain score approach within each trajectory subgroup, compared to the full sample data set. Using the imputed data, the results (Table 11) indicated that all the results are inconsistent from using the residual gain score approach and using the simple gain score approach within each subgroup. In the low or none depression group, the psychological treatment effect sized are $d_I = -0.02$ ($p > 0.1$) using the simple gain score approach and $b_I = 1.63$ ($p < 0.001$) using the residual gain score approach and medication effect are $d_I = -0.18$ ($p >$

0.1) using the simple gain score approach and $b_I = 1.32$ ($p < 0.01$) using the residual gain score approach. In the High depression group, psychological treatment effect are $d_I = -1.70$ ($p < 0.001$) using the simple gain score approach and $b_I = 1.31$ ($p < 0.001$) using the residual gain score approach and medication effect are $d_I = -1.64$ ($p < 0.001$) using the simple gain score approach and $b_I = 1.12$ ($p < 0.01$) using the residual gain score approach. In the W-2 depression group, psychological treatment effect are $d_I = -1.26$ ($p > 0.1$) using the simple gain score approach and $b_I = 1.53$ and $p < 0.1$ using the residual gain score approach and medication effect are $d_I = 1.59$ ($p > 0.1$) using the simple gain score approach and $b_I = 3.30$ ($p < 0.001$) using the residual gain score approach. However, as expected and similar to the simulated data results, consistent results are reached from the group-centered ANCOVA approach that all result from the simple gain score approach and the residual gain score approach using the centered data are consistent to the results from the simple gain score approach using imputed unmated data. In the low or none depression group, psychological treatment effect are $d_I = b_I = -0.02$ ($p > 0.1$) and medication effect are $d_I = b_I = -0.18$ ($p > 0.1$). In the High depression group, psychological treatment effect are $d_I = b_I = -1.70$ ($p < 0.001$) and medication effect are $d_I = b_I = -1.64$ ($p < 0.001$). In the W-2 depression group, psychological treatment effect are $d_I = b_I = -1.26$ ($p > 0.1$) and medication effect are $d_I = b_I = 1.59$ ($p > 0.1$).

Research Question 4

The previous step indicated that group-centered ANCOVA is an easy way to produce consistent results across both gain score approaches, but it is an artificial way of getting consistent results. Its results are unbiased only if the simple gain score approach

gives unbiased results, according to the simulation study from the previous chapter. However, in the FFCW data, there is no way to tell whether the consistent result is biased or not. In addition, consistent results could not be reached except when using the Centered ANCOVA approach in the previous step. Therefore the next step is to investigate whether the model can be improved by answering Research Question # 4. That Research Question asked whether employing propensity score adjustment approach can get more convincing evidence of approximating unbiased causal evidence. The analysis is a combination of propensity score adjustments and mixture modeling, which produced consistency across those two baseline adjustment methods in Haviland and colleagues (2007). In order to compare the results from the simulation study, I first ran a condition in which the propensity scores were based on the pretest outcome ONLY (depression at wave 4) within each trajectory subgroup. Then I generated the usual type of propensity scores based on the 16 most relevant covariates within each trajectory subgroup. Propensity in the logit for each covariates are in Table 12.

After generating propensity scores and before matching samples, I tested for sample balance between the active treatment group and the comparison group within each homogeneous subgroup, following the example of Haviland and colleagues (2007). The boxplots in Figure 14 show the propensity scores calculated from the pretest score ONLY. The boxplots in Figure 15 show the propensity scores calculated from all 16 relevant covariates. Both figures compare the range of propensity scores for the active treatment group and the comparison group for the total *unmatched samples* within each trajectory subgroup. The two figures show that propensity scores in the active treatment group and the comparison group have some overlapping cases within each trajectory

subgroup for each treatment condition, indicating that it is possible to get matching samples from the comparison group and the active treatment group.

Then, the propensity scores prior to treatment are used for matching samples from the comparison group to the active treatment within each sub-group. When propensity scores are based on the pretest only, it is possible to get exact matches on propensity scores, which is then equivalent to matching on pretest scores, as shown in Figure 14. When propensity scores are based on many covariates, however, it is difficult to get exact matches (Figure 15). Consequently propensity score matching selects a caliper to specify how closely two matched cases need to be to make a match. For each matching in this study, the calipers varied from 0.2 to 0.5 and the matching ratio for the relative sizes of the control samples matched to treated samples varied from 1:1, 2:1, 3:1, to 4:1. Initially, the caliper was set to 0.2 and the matching ratio was set to 4:1. If the matching was not successful, the matching ratio was decreased one step at a time until the matching succeeded. If reducing the matching ratio to 1:1 could not achieve successful matching, the caliper was increased 0.1 each time until the matching was successful. The boxplots in Figure 16 show the propensity scores based on the pretest score ONLY, and the boxplots in Figure 17 show the propensity scores based on multiple covariates, comparing the propensity scores for the matched active treatment group and the matched comparison group within each trajectory subgroup. The two figures show that samples are successfully matched within each trajectory subgroup based on the available covariates.

After matching, the balance for each covariate was checked. Table 13 shows the balance condition of each covariate comparing before matching and after matching for

each trajectory subgroup when the propensity scores are calculated based on multiple covariates. Table 13 shows the standardized difference (Cohen's d) before and after matching within each subgroup. Using the cut point of 0.2, Table 13 indicates that in the High Depression group, after matching, all the covariates are balanced for both the psychological treatment and the medication treatment. In the Low Depression group, mothers' depression at Wave 2 and domestic violence are not balanced between the psychological treatment group and the comparison group and mothers' depression at Wave 3, financial hardship, domestic violence, and parenting stress are not balanced between the medication treatment group and the comparison group. In the Medium Depression group, half of the covariates are not balanced between the psychological treatment group and the comparison group and mothers depression at Wave 2 and social support are not balanced between the medication treatment group and the comparison group. In addition, three variables have only zero scores in a comparison group: mothers drug use for medication treatment in the Low depression group, and mothers foreign born for the comparison groups for both kinds of treatment in the Medium depression subgroup. The sample balance check after matching indicated that only the High depression subgroup have completely matched samples on propensity scores based on the 16 covariates.

Next, since the assumption of normality and the assumption of homogeneous pretest scores are still violated in the three subgroups, these two assumptions were diagnosed using the matched data within each subgroup. The distribution of mothers' depression at Wave 5 using matched samples are in Figure 19 (a), (b), and (c), based on using the pretest as the only covariate and in Figure 20 (d), (e), and (f), based on

propensity score calculated from the 16 true confounders. Figure 18 showed that at least 35% of the matched samples reported zero depression scores at Wave 5 in each subgroup, indicating the distribution of the outcomes within each subgroup are still not normal. The distribution of residuals are in Figure 19, based on propensity scores calculated from pretest scores only, and in Figure 20, based on propensity scores calculated from 16 covariates. Figure 19 indicated that when propensity scores are calculated from the pretest only These distributions don't look like normal distributions, although some of them are less skewed than other and/or are less dominated by one peak value. In addition, pretest mean differences are tested. The results in Table 11 show that using the matched sample including the mothers with no depression symptoms at Wave 5, the pretest means are not significantly different between the active treatment group and the comparison group. Therefore, only the normality assumption is violated within each matched subgroup. Even though the normality assumption is violated, the results from the simple gain score approach and the residual gain score approach consistency on their direction, and their effect sizes get closer to each other. However, the results are mostly in the direction of the original ANCOVA results, indicating that treatment is harmful by predicting higher depression severity, significantly so for most analyses in the High Depression group, which has the larger sample and greater statistical power of the two groups with moderate-to-high depression severity at Wave 4.

Since the normality assumption was still violated with matched samples within each trajectory subgroup, I excluded cases with zero depression scores at either Wave 4 or Wave 5 (the truncated sample) in the matched samples within the three trajectory subgroups, and re-ran the analysis within each subgroup. I tested the pretest group mean

difference using the truncated matched sample. The results showed that there is no significant difference on pretest group means between the treatment group and the comparison group within each trajectory subgroup. Then I tested whether I can get less biased results using the simple gain score approach and the residual gain score approach using the truncated matched sample within each subgroup. The results indicated that all the results are consistent in predicting harmful effects of treatment (occasionally significantly). In addition, only two analyses, based on the truncated matched on multiple covariate data, show that psychological treatment in the low depression subgroup and medication treatment in the medium depression subgroup may benefit by reducing depression severity, but neither effect is significant and all other analyses make psychological treatment or medication treatment appear to be harmful in increasing depression severity, either significantly or not significantly. However, the matching balance check in Table 14 indicated that the low depression group and the medium depression groups did not match well since some covariates remain unbalanced as indicated by standardized differences larger than 0.2. Thus, the combination of the latent class models and propensity-score matching can achieve consistent direction of the results, but the consistent results indicate harmful average effects of both treatments for depression, which looks significantly harmful for the largest subgroup with the highest mean depression severity at Wave 4. This indicates either that both of these common treatments for depression are harmful on average for the most depressed low-income mothers in large American cities, or that these “best” results are biased against corrective action type of treatment (Larzelere, Lin, Payton, Washburn, in press), consistent with the simulation data.

Discussion

The purpose of the second study using the FFWC data is to apply lessons learned from the simulated data to analyses of real data, to examine whether I can get consistent and less biased results from the simple gain score approach and the residual gain score approach by improving the models. Four research questions were addressed in this chapter to confirm contradictory results for analyses of the two types of gain scores, diagnose violations of ANCOVA assumptions, minimize or adjust for violations of assumptions, and apply group-centered ANCOVA approach and Haviland and colleagues' (2007) combination of mixture modeling and propensity score matching to test when the two approaches give consistent results and whether those results are less biased than the standard models.

As expected, two-wave data analyses and the three-wave data analyses produce contradictory results between the simple gain score approach (e.g., the latent growth model across 3 waves) and the residual gain score approach (the cross-legged panel model across 3 waves). Then assumption diagnoses showed that the assumptions of normality, homogeneous slope (only for psychological treatment), and homogeneous pretest group means were violated in the FFCW data for the two types of treatment for mothers' depression.

Next, linear latent class models using a negative binomial hurdle model were used to test whether the data are composed of heterogeneous subgroups using mothers' depression data at Wave 2, Wave 3, and Wave 4. The test results identified three more homogeneous trajectory subgroups. Within each trajectory subgroup, I tested whether I

could get consistent results from the simple gain score approach and the residual gain score approach, and tested whether the assumption of normality, homogeneous slopes, and homogeneous pretest group means were still violated within each trajectory subgroup. Similar to the simulation study, although slopes were now homogeneous within the subgroups, consistent results could not be achieved because pretest means for the treatment group and the comparison group were still significantly different, and the distribution of the outcome residuals were not normally distributed. This finding in the FFCW data supports the results from the simulated data in that, although other assumptions of ANCOVA and simple gain score approach are met, the results remain inconsistent if the pretest means differ significantly between the treatment group and the control group.

When using either group-centered ANCOVA approach or the combination of latent class models and propensity score matching to minimize pretest differences, the results from the simple gain score and the residual gain score become consistent on the direction of the apparent effect. When using these two approaches, the pretest means are not significantly different anymore but the normality assumption is still violated. This supports the results from the simulation study that once the pretest difference is removed, the results will be consistent even if other assumptions are still violated. Similar to the simulation study, the results are in the same direction as the simple gain score approach when using the group-centered ANCOVA approach, whereas the results are in the same direction as the residual gain score approach (except for two non-significant results using truncated samples) when using the combination of latent class models and propensity score matching. However, unlike the simulated data where we know which null

hypothesis is true and can thereby tell whether a statistical approach is biased, we have no way to know with certainty which approach is less biased using the FFCW data.

Compared to results using matching based on propensity scores calculated from using pretest scores only (depression at Wave 4), results matching based propensity scores calculated from the 16 true confounders remains the same direction in treatment effect, but the effect size and the significant levels may vary.

When using the truncated sample, the problem of violating the assumption of normality is removed, but balance between the treatment group and the matching control group is not achieved in the low depression and medium depression subgroups. The reason is that within these two subgroups, some covariates either only have zero scores (drug used for medication treatment within the low depression subgroup, and mother foreign born for psychological treatment and medication treatment within the medium depression subgroup) in the comparison group or the standardized difference between the treatment group and comparison group are greater than 0.20. Within the high depression subgroup, which is matching well whether cases with Wave 4 or Wave 5 zero scores are excluded or not, similar to the simulation studies, the results indicate that both treatments appear to be harmful by increasing depression severity either significantly or not significantly.

CHAPTER V

GENERAL DISCUSSION

Summary and Interpretation of Results

The major purposes of human development study are to describe and to explain human change (within-person change and between-person differences in those changes) in order to promote improvements in human life. One foundation of promoting improved human life is accurate explanation, which relies on making valid causal inference. Making valid causal inference has stricter requirements for analyses than simple prediction, involving the three criteria (correlation, sequence, and unique explanation) for valid causal inference. The first two criteria, correlation between the cause and the effect and the temporal sequence of the cause and the effect, could be met through research design. The third criterion, a unique explanation, requires there to be no plausible alternative interpretation of the temporal association between the cause and the effect other than the treatment, namely no confounder effects. In human development studies, confounding variables are common, and the common statistical method for reducing confounding effects is to include confounders as covariates in the analysis.

Finally, the logic of a causal relationship, as reflected in the counterfactual model, compares outcomes after being treated and not being treated at the same time within the same person (or unit), which is impossible in reality. This has been called the fundamental problem of causal inference, namely that it is impossible to know the difference in any outcome that would have occurred if a person would have selected the opposite treatment condition than they did. However, Rubin's Causal Model uses that definition to clarify the difficulty to be overcome to make unbiased causal inferences for the average effect, even if that effect could vary from one person to another person. The central implication is that the treated and non-treated groups must be made equivalent in their average pre-treatment prognosis on the outcome variable, so that any differential outcomes can be attributed to differences in the effects of the treatment and comparison groups. The requirement of no prior significant difference in prognosis between the treated units and the comparison units guarantees that the two studied groups of units are comparable. This requirement could be met in randomized studies that randomly assign samples to be in the treatment condition or in the comparison condition, which make experimental units in different treatment conditions to be similar on average and therefore comparable. However, for ethical or practical reasons, randomized studies can rarely be manipulated in human development studies. This results in human development research usually involving self-selection that creates known or unknown differences between the treatment group and the comparison group before the treatment conditions are introduced, so that the two experimental groups are not comparable. In order to solve this problem, human development investigators use a simple gain score approach or a residual gain score approach to "control" for pretest differences. However, "controlling for" the pretest

difference, using ANCOVA or the simple gain score approach, could lead to inconsistent results between the two approaches, which is known as Lord's paradox. Lord's paradox also provides a signal that different statistical methods could produce inconsistent or even contradictory results, which raises questions about which statistical method is less biased for making causal inference. Lord's (1967) paradox has been discussed for more than 50 years. The consensus is that the paradox does not exist in randomized studies using the simple gain score approach and the residual gain score approach (Van Breukelen, 2013). However, it is not clear what leads to the inconsistent results between the simple gain score approach and the residual gain score approach and whether it is possible to get consistent results from the two approaches in nonrandomized studies. The purpose of the current study is to understand the mechanism of selection bias in nonrandomized studies by exploring possible reasons for inconsistent and biased results between the simple gain score approach and the residual gain score approach. Specifically, I address Lord's (1967) paradox, diagnose possible reasons that lead to the paradox, and test alternative models for solving the problems of the paradox and for reaching consistent and less biased results using simulated data and using real data on treatment for mothers' depression from the Fragile Families and Child Wellbeing (FFCW) data.

The results indicated that: 1) Lord's (1967) paradox exists when the mean pretest scores are different for the treatment group and the comparison group, and consistent results can be reached when the mean pretest difference between treatment group and comparison group is removed. 2) Group-centered ANCOVA approach and the combination of mixture modeling and matching on either the pretest or the propensity score could be used to remove the pretest difference between the treatment group and

comparison group. 3) The latter two methods of removing the mean pretest difference between the two treatment conditions result in two distinct sets of consistent results, but those consistent pairs of results are nearly as inconsistent with each other as the original inconsistency. 4) Consistent results do not guarantee an unbiased result.

First, I tested Lord's paradox using the simulated two-wave data and the treatment for mothers' depression in the FFCW data in two-wave and three-wave analyses. In the simulation studies, I generated two sets of data: the first to fit the null hypothesis of simple gain scores, consistent with the original figure in Lord's (1967) paradox. The second simulated data was designed to fit the null hypothesis of ANCOVA, which I called reversed Lord's paradox. It assumes that no group difference in weight gains is represented by the amount of shrinkage between the group means that is estimated by the within-group slope from pretest to posttest. Generation of the two data sets allowed me to test which approach is biased between the simple gain score approach and the residual gain score approach. The results indicated that when the simulated data fit the null hypothesis of the simple gain score approach, the result is unbiased using the simple gain score approach and biased using the residual gain score approach. On the other hand, when the data fit the null hypothesis of ANCOVA, the results are biased using the simple gain score approach and unbiased using the residual gain score approach. This suggests that either approach can give a correct answer, if its null hypothesis is an unbiased estimate of a zero treatment (group) effect.

In the FFCW data study, as expected, similar to previous studies (Berger et al., 2009; Larzelere et al., 2010) and the simulations, the results showed that contradictory results exist between the simple gain score approach and the residual gain score approach

when using two-wave analyses. The contradictory results also exist between the cross-lagged panel model and the latent growth model when analyzing three waves of data. The paradox in three-wave analyses suggests that contradictory results could exist in complex statistical modeling as well as in simple two-wave modeling. Both the two-wave data and the three-wave data show that treatment helps to reduce depression symptoms according to the simple gain score approach (latent growth model) but treatment increases depression symptoms according to the residual gain score approach (cross-lagged panel model). Unlike simulated data, however, it is not clear in analyses of real data which results will be biased and which results will be unbiased.

One possibility is that Lord's paradox occurs when the assumptions of one or both types of analyses are violated. Therefore, the consequences of violations of the assumptions of ANCOVA and simple gain scores were diagnosed (i.e., normality of outcome residuals, homogeneous slopes within the treatment and the comparison groups, and independence of treatment and covariates). The results indicated that inconsistent results occur mainly due to violating the assumption of independence between the covariate and the treatment. The results from the simulation study (Table 2) indicated that once the assumption of independence between covariate and treatment is violated, violations of other assumptions further distort the results of ANCOVA, which supports previous studies by Levy (1980) and by Sullivan and D'Agostino (2002). When the assumption of independence between covariate and treatment is not violated, violation of other assumptions did not lead to the paradox. These results are further supported by the results from the second study in the treatment for mothers' depression in the FFCW data, which indicated that contradictory results remain when assumptions other than the

assumption of independence between treatment and covariate are met. Moreover, when the pretest mean group difference is removed, the results from the simple gain score approach and the residual gain score approach are consistent and very close to each other.

In order to determine how much the pretest difference will affect the difference of the effect size estimates between the simple gain score approach and the residual gain score approach, in the simulation study, I simulated data to keep the males' pretest weight and posttest weight the same and varied the females' pretest and posttest weights. The results indicated that the effect size from the two approaches would not be the same unless there is no difference between the group mean pretest scores. Results from the residual gain score approach are biased in the direction of the pretest mean group difference compared to the simple gain score approach. The difference of the effect sizes between the two approaches is determined by the pretest difference between the treatment group and the comparison group and the within-group slope coefficient in the residual gain score approach: $d - b = (1 - slope)(\bar{Y}_{20} - \bar{Y}_{10})$, where d is the effect size using the simple gain score approach, b is the effect size using the residual gain score approach, $slope$ is the slope coefficient predicting the posttest from the pretest in the residual gain score formula, and $(\bar{Y}_{20} - \bar{Y}_{10})$ is the pretest mean difference between the treatment group and the comparison group. This is consistent with previous comparisons of the two approaches, which typically state that the simple gain score approach requires the slope to be 1.00, with the implication that it is better to estimate the slope from the data. However, ANCOVA could be regarded as just as rigid as the simple gain score approach, in that the estimated shrinkage of the distance between the group means on the outcome variable from pretest to posttest is expected to be exactly estimated by the within-group slope

coefficient, according to ANCOVA's null hypothesis. In contrast, the simple gain score approach assumes that the expected shrinkage between the group means is not estimated by the within-group slope, but it incorporates its own rigid estimate of that (lack of) shrinkage for its null hypothesis.

Given that a pretest mean difference between the treatment group and the comparison group will lead to inconsistent results between the simple gain score approach and the residual gain score approach, would that inconsistency be resolved by removing the pretest group difference? If so, how should the group difference on the pretest be removed? First, I developed the group-centered ANCOVA approach that centers both the pretest and the posttest scores around the pretest group means. By centering the pretest scores around their group means, the difference in the pretest group means becomes zero, and the pretest group differences are removed. By centering the posttest scores around the pretest group means, the change from the pretest to the posttest scores remained unchanged for each person. This was adapted from Huitema's (2011) quasi-ANCOVA, which controls for a post-treatment covariate by centering its scores around its group means. As expected, by using the group-centered ANCOVA method, the results from the simple gain score approach and the residual gain score approach are consistent. However, their effect sizes were exactly the same as the estimated effect size from the simple gain score approach. The group-centered ANCOVA provides greater power than simple gain score analyses and therefore may be preferable when the latter provides an unbiased or less biased causal estimate.

Second, I implemented a simple and more complex version that combined mixture modeling and propensity score matching, following Haviland et al. (2007). The

simple version adjusted only for the pretest score in the simulation study and the FFCW study, whereas the more complex version used propensity scores based on 16 covariates in the FFCW data. It was expected that when more covariates were added to the analysis, the “harmful” effect size for treating depression would be reduced, if not reversed. The results indicate that the “harmful” effect size was reduced in the low depression and medium depression group but not in the high depression group when controlling for additional covariates to calculate propensity scores, compared to matching only on pretest scores. When matching propensity scores calculated based only on the pretest score, results from the matched data are very close to results from the residual gain score approach using the unmatched samples. One possible interpretation is that propensity score adjustments are only as good as the covariates that are used to calculate them (Steiner, Cook, Shadish, & Clark, 2010). If the covariates are good enough to reduce pretreatment confounding, the results should be less biased.

Moreover, since the group-centered ANCOVA approach statistically removed the pretest difference artificially and did not actually remove the effect of the pretest difference on treatment to get unbiased results, it was expected that the combination of the mixture modeling and the propensity score approach could provide consistent and less biased results than the group-centered ANCOVA approach. It is surprising that although consistent results can be reached using the combination of mixture modeling and the propensity score method, those consistent results did not guarantee less biased results. The evidence from the simulation studies indicated that the consistent results from simplistic versions of mixture modeling (above and below the grand mean pretest weight) and propensity score matching (matching on the pretest only) produced unbiased results

only when the simulated data fit the null hypothesis on ANCOVA. The results for mixture modeling and propensity score matching on treatment for mothers' depression using the FFCW data showed that the two treatments are either significantly or not significantly harmful for depression, although the results using matching data have smaller effect sizes than the results from the ANCOVA approach using unmatched subgroup samples or the overall sample. These results are opposite to the findings in randomized studies or meta-analysis (Andersson et al., 2009; Cuijpers, et al., 2006; Muet. al., 2006; Parekh, 2017) indicating that treatments for depression either have no effect or help in reducing depression symptoms, indicating that bias still remains in the current study even using the combination of mixture modeling and the propensity score method.

Model Comparisons

Overall, both results from the group-centered ANCOVA approach and the combination of the mixture modeling and the propensity score method made the groups have equal pretest means, which produced consistent results. However, both the simulation study and the depression study in the FFCW data indicated that the group-centered ANCOVA approach will bring the consistent results toward the direction of the simple gain score approach and the combination of the mixture modeling and the propensity score method will bring the results toward the direction of the ANCOVA approach. These new effect sizes from the two consistent results are not consistently closer to each other than the distance between the original simple vs. residual gain score approaches. So, which type of analysis may be less biased? Are complex models less biased than simple models? What can we learn from the current study?

ANCOVA vs. Propensity Score (Combining or not Combining Mixture Modeling)

The problem of ANCOVA for “reducing” selection bias in nonrandomized studies is its violation of the assumption of independence between treatment and the covariates. Consequently, it controls for the confounding effect of the pretest difference on the outcome, but fails to eliminate selection bias. That being the case, does a propensity score method minimize the problem of selection bias by matching samples on their propensity score to reduce the mean difference of pretest scores between the treatment group and the comparison group? Rubin thinks it is sufficient to match on propensity scores. Campbell thinks that regression toward different subgroup means can produce differences in prognoses even for matched cases. In the simulation data, I tested their differential predictions by creating matched pairs from groups that differ on their average regression toward their own group mean. I got results similar to ANCOVA with propensity score matching or matching on the pretest, which supports Campbell more than Rubin. Matching on the pretest or on propensity scores does not necessarily match the groups on their pre-treatment prognoses. The propensity score approach is similar to an ANCOVA approach with the same set of covariates, according to the current study. This also supports Steiner and colleagues’ (2010) conclusion that having the right covariates is more important than whether one used residual gain score analyses or propensity score methods to adjust for those covariates.

Does a combination of mixture modeling and propensity score matching improve the results? To answer that question, I combined mixture modeling and propensity score matching in the analyses. I first identified trajectory groups using Latent Class Analysis using a linear regression model to get trajectory subgroups and tried to match samples

based on propensity scores within each subgroup, but matching was hard to achieve. Considering the distribution of depression was zero-inflated, I then retested trajectory groups by rerunning a Latent Class Analysis using a Negative Binomial Hurdle Model to get trajectory subgroups and tried to match samples based on propensity scores within each subgroup, matching on propensity scores generated by the 16 covariates was achieved for each subgroup. However, the samples were balanced on all 16 covariates only in the high depression subgroup. In the medium depression and the low depression subgroups, at least two of the covariates were not balanced. This situation also happened in Haviland and colleagues' (2007) study. When they combined their mixture modeling and propensity scores, they dropped one subgroup in which propensity scores failed to overlap sufficiently for matching. This shows a drawback of the combination of mixture modeling and propensity scores, that matching can be hard to achieve when the sample sizes get smaller within each trajectory subgroup even though the analyses were conducted in a large sample size database such as the FFCW data with 4898 mothers. The current study suggests that a combination of mixture modeling and propensity score methods are not superior to a residual gain score approach that controls for the same covariates.

Simple Gain Score vs. Residual Gain Score

The second question is whether using the residual gain score approach can produce less biased result than using the simple gain score approach? Results from the simulation and the treatment for depression using the FFCW data indicated that the inconsistent results occurred mainly due to pretest difference. When pretest group mean scores differ, the difference of the effect sizes between the simple gain score approach

and the residual gain score approach is $b_1 - d = (1 - slope)(\bar{Y}_{20} - \bar{Y}_{10})$. The right side of the formula, $(1 - slope)(\bar{Y}_{20} - \bar{Y}_{10})$, tells us that the difference in results from the two approaches is related to the difference in the group means on the pretest ($\bar{Y}_{20} - \bar{Y}_{10}$) and the *slope* of the posttest score regressed on the pretest score in the residual gain score formula. This also suggests that violation of the assumption of independence between covariates and the treatment is the essential factor that contributes to the paradox. Note that the assumption of independence between the covariates and the treatment is one of the assumptions of ANCOVA but not an assumption of simple gain score approach, which suggests that biased results are mainly due to the ANCOVA approach when pretest mean scores differ between the treatment group and the comparison group.

Moreover, recall that human development studies are usually based on some combination of within-person change and between-person differences. These two summary statistics seem to contribute to the simple gain score approach and the residual gain score approach in different ways (see Figure 21). The residual gain score approach seems to combine between-person differences that are due to within-person changes and between-person differences that are not due to within-person changes (Berry & Willoughby, 2017; Hamaker et al., 2015; Hoffman, 2015). In the left of the Figure 21, we can see the residual gain score approach is used to explain the remaining variance of posttest between-person differences, which is the dependent variable. The variance of the posttest between-person differences could be due to the treatment effect resulting in within-person changes and other factors including pretreatment between-person differences and confounding effects after treatment. The effect of other known factors could be partialled out when including the confounders or the pretreatment conditions as

covariates. The problem is when the pretest scores are different between the treatment group and the comparison group, the pretest differences between the treatment and comparison group will influence the treatment assignment and further influence the outcome (the dashed-line path in Figure 21). In other words, pretreatment differences not only directly influence the outcome (posttest between-person differences) but also indirectly influence the outcome through the treatment assignment. Thus, the estimated treatment effect from the residual gain score approach is a smushed (Hoffman, 2015) treatment effect, including the pure within-person treatment effect and the treatment effect that is due to between-person differences affecting treatment assignment. The smushed treatment effect is biased in analyzing corrective action types of treatment (Larzelere, Lin, Payton, Washburn, in press), i.e., treatments expected to reduce the differences of outcome scores between the treatment group and the comparison group. It is also biased in analyzing exaggerating types of treatment, treatments expected to increase the differences of outcome scores, but in a different way. Examples for corrective action types of treatment include marriage therapy for reducing the propensity of divorce, Head Start for improving school performance, and parental divorce for decreasing children's willingness to get married. Examples for exaggerating types of treatment include joining a gang, which increases delinquency and marital hostility, which increases the propensity of divorce. When analyzing corrective action types of treatment, since the pretest group mean score is positively correlated with the posttest score (perceived likelihood of divorce) and the treatment is attempting to reduce that outcome (the posttest score) gap, the estimated treatment effect is the result of the contribution of the pure treatment effect counterbalancing the contribution of the pretest

difference confounded with treatment assignment. If the confounded effect of the pretest differences correlated with treatment assignment on the outcome is much stronger than the pure treatment effect, the estimated smushed treatment effect will be positive, suggesting that marriage therapy is associated with a higher probability of divorce. If the effect of the pretest differences on the outcome that is confounded with treatment assignment is much weaker than the pure treatment effect, the estimated smushed treatment effect will be negative, i.e., beneficial in reducing the desire to divorce. Whether the effect of pretest differences confounded with treatment assignment on the outcome is stronger or weaker than the pure treatment effect, the estimated smushed treatment effect will always be smaller than the true treatment effect in corrective action studies. For example, marital therapy is a corrective action for decreasing the propensity of divorce. Attending or not attending marital therapy is influenced by the pretest of the propensity of divorce, which will result in an estimated treatment effect of the marital therapy as a smushed treatment effect in non-randomized studies. Since attending marital therapy is negatively correlated with the posttest of the propensity of divorce and the pretest of the propensity of divorce is positively correlated to the posttest of the propensity of divorce, the smushed treatment effect is the pure treatment effect of the marital therapy counterbalancing the effect of the pretest group differences in the propensity of divorce. Whether the estimated treatment effect (smushed) of the marital therapy is positive or negative depends on which side of this counterbalancing is stronger. When the pretest of the propensity of divorce between the treatment group and the comparison group are significantly different and these differences strongly influence whether a couple goes to the marital therapy or not and further influences the outcome

more strongly than the counterbalancing of the pure treatment effect on the outcome, the estimated treatment effect would be positive indicating treatment is harmful. This could be used to interpret why treatment appears to be harmful for depression in the current study, why non-physical punishment appears to be harmful for child outcomes (Larzelere et. al., 2010), and why Summer Head Start program appeared to have negative effects on child well-being using the residual gain score approach. In contrast, when analyzing exaggerating actions, a type of treatment that increases pre-existing group differences on the outcome, since both pretest group mean differences and treatment are positively correlated with posttest scores, the estimated treatment effect (smushed) is the pure treatment effect plus the confounding effect of the pretest difference on treatment assignment, which further affects the outcome. The smushed treatment effect will always be positive and larger than the pure treatment effect. Thus, the residual gain score approach could produce either direction of biased results when the pretest scores are different between the treatment group and the comparison group.

One the other hand, the simple gain score approach, which analyzes how within-person changes across two time points can lead to between-person differences regarding change, emphasizes that the accumulation of within-person changes is the reason for between-person differences at any given time (see the left side of Figure 21). That approach therefore ignores between-person difference that are not due to within-person changes in the period studied. When there is no pretest difference between the treatment group and the comparison group, the comparison of changes between the two groups will be expected due to the treatment effect, and it is expected that the comparison group will not have any treatment effect. This seems to reflect the counterfactual model for inferring

a causal effect in that the counterfactual would have no treatment effect since it did not receive treatment. In this perspective, the simple gain score approach is more close to the definition of making casual inference and seems less biased. In addition, in randomized studies, since the between-person difference that is not due to within-person change is removed by random assignment, the between-person difference ONLY counts the part that is due to within-person change in the residual gain score approach. That is why the results from the residual gain score approach is consistent with the simple gain score approach. At this point, we can see, the main purpose of a randomized study is to remove the between-person difference that is not due to new within-person change and only to estimate the part of the posttest between-person difference that is due to new within-person change. Thus, statistical models based on the simple gain score approach, including the latent growth model, seems superior to models based on the residual gain score approach, including the cross-lagged panel model, for making casual inference, based on the results from the current study.

Simple Gain Score vs. Group-Centered ANCOVA

Although statistical models based on the simple gain score approach may be superior to models based on the residual gain score approach for making casual inference, it does not mean that the simple gain score approach is unbiased in nonrandomized studies. One drawback of the simple gain score approach is that it has less power than the residual gain score approach. In addition, one assumption of the simple gain score approach is that any between-person differences in within-person change is 100% due to the treatment effect, which is impossible in most human development studies where confounding effects are common. Under this assumption, any other contributions to

within-person change will count as an effect of the treatment even if there is no treatment effect. Thus, the estimated treatment effect could be biased using the simple gain score approach when confounders' effects exist but could not be partialled out. This two drawback may be overcome using complex models based on simple gain score such as the latent growth model by including confounders in the model.

Alternatively and simply, the results in this study suggest that a group-centered ANCOVA approach could also overcome a major drawback of the simple gain score approach. When using the group-mean centered data, the result from the residual gain score approach has more statistical power than the simple gain score approach. Also, using the simple gain score approach, the reason that leads to within-person change should ONLY be treatment. Otherwise, any other contribution that is correlated with treatment conditions and leads to differential within-person change will be credited to the treatment and will result in biased results. The residual gain score approach using the group-mean centered data should be able to control for other factors that lead to change. For this reason, the residual gain score approach is better than the simple gain score approach using group-mean centered data.

Note that the current study does not show a combination of mixture modeling and the group-centered ANCOVA approach is less biased than the simple version of group-centered ANCOVA. However, the current study does indicate that when using group centered-ANCOVA, the assumption of normality of residuals always needs to be checked and statistical models should fit the distribution of the data, as necessary.

Directions for Human Development Studies

In sum, the current study suggests for human development researchers, if the research interest is on comparing between-person differences such as comparative studies on global families without considering time-related changes, models based on residual gain scores that focusing on explaining the variance of between-person differences should be considered. When using models based on residual gain scores, the assumption of independence of covariate and treatment and the assumption of normality of outcome residuals should always been checked first. When these two assumptons are violated, other assumptions such as homogeneous slope and homogeneous variance between the treatment group and the comparison group should be further examined. If the assumption of independence of covariate and treatment is violated, researchers should note that the results may be biased only in magnitude of the effect size if the analysis is on an exaggerating type of treatment whereas the results may be biased in both the direction and the effect size if the analysis is on a corrective action type of treatment. On the other hand, for human development researchers, if the research interest is on making casual inference about within-person change such as whether the Head Start progrom reduces children's antisocial behavriors or whether marital therapy reduces marital conflilct, models based on simple gain score that focus on explaining the pattern of within-person change should be considered. When using models based on simple gain scores, models such as group-centered ANCOVA for two-wave data analyses and latent growth models for more than two waves are superior since these types of models have more statistical power and are able to control other confounding effects, compared to the two-wave simple gain score approach.

Strengths

This study has several important strengths. The greatest strength is that it uses Lord's paradox to understand how the simple gain score approach (latent growth model) and the residual gain score approach (cross-lagged panel model) contribute to analyses of the two types of human development changes (within-person change and between-person differences). Specifically, the simple gain score approach focuses on how accumulated within-person change leads to new between-person differences and ignores the analysis of between-person difference that not due to new within-person change. On the other hand, the residual gain score approach combines between-person differences that are due to new within-person changes and between-person differences that are not due to new within-person changes (Berry & Willoughby, 2017; Hamaker et al., 2015; Hoffman, 2015). This study provides new guidelines to help human development researchers choose among statistic methods for making causal inference in non-randomized studies. Second, this is the first known study that uses simulated data to fit the null hypothesis of the simple gain score approach and to fit the null hypothesis of ANCOVA, which allowed me to examine which result is biased and which result is unbiased. It also contributed to the goal to help human development studies by revealing how the inconsistent results from the simple gain score approach and the residual gain score approach can be explained to help clarify their specific meanings and appropriate applications. In addition, in order to understand the reasons for Lord's paradox, it conducted multiple diagnoses on the possible contributions that lead to Lord's (1967) paradox. I tested the results due to violating assumptions of ANCOVA and the simple gain score approach and found that violation of the assumption of independence of

treatment and the covariates is the key that contributes to Lord's paradox. It is also the first empirical study to introduce the group-centered ANCOVA approach. Moreover, this study found that consistent results from the simple gain score approach and the residual gain score approach do not guarantee the results are unbiased. Finally, this study used both simulated data and real data to conduct the analyses, which strengthen the evidence of results on the paradox diagnoses and the model comparisons.

Limitations and Future Suggestions

Although there are several strengths in this study, it is important to point out the limitations. According to the temporal sequence requirement, a cause must happen before the effect. Accordingly I excluded samples that had already received treatment for depression at Wave 3, but that does not guarantee that mothers did not receive treatment at Wave 2 or Wave 1 or even earlier. This limitation can produce biased results. In addition, the pretest scores (depression severity at Wave 4) are measured at the same time as the two types of treatment at Wave 4, so that pretest depression may already be affected by the treatment at Wave 4, which may also provide biased results. Another limitation of this study is that confounding effects on the change of depression after the treatments were introduced have not been controlled. As mentioned in the first paragraph of this chapter, confounders besides the treatment could affect the change of the outcome at the same time when the treatments were introduced and have their effect before the posttest outcomes were measured. In the present study, my focus was on reducing selection bias that happens before the treatment was introduced, and I did not control for time-varying confounders affecting subsequent changes in the outcome. Future research

should continue to explore controlling the confounding effects that happen at the same time and work with the treatment to affect the outcome.

REFERENCES

- Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology*, 20, 93–114. doi.org/10.2307/271083
- Andersson, G., & Cuijpers, P. (2009). Internet-based and other computerized psychological treatments for adult depression: a meta-analysis. *Cognitive Behaviour Therapy*, 38, 196-205. doi:10.1080/16506070903318960
- Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's approach*. Princeton University Press: Princeton, NJ.
- Beck, C. T. (1996). A meta-analysis of the relationship between postpartum depression and infant temperament. *Nursing Research*, 45, 225-230.
- Berger, L. M., Bruch, S. K., Johnson, E. I., James, S., & Rubin, D. (2009). Estimating the “impact” of out-of-home placement on child well-being: Approaching the problem of selection bias. *Child Development*, 80, 1856-1876. doi.org/10.1111/j.1467-8624.2009.01372.x
- Berry, D., & Willoughby, M. T. (2017). On the practical interpretability of cross-lagged panel models: Rethinking a developmental workhorse. *Child Development*, 88, 1186-1206. doi.org/10.1111/cdev.12660

- Blumberg, C. J., & Porter, A. C. (1983). Analyzing quasi-experiments: Some implication of assuming continuous growth models. *Journal of Experimental Education*, 51, 150-159. doi.org/10.1080/00220973.1983.11011854
- Brorsen B. W. (personal communication, March 9, 2018)
- Butler, S. M., Black, D. R., Blue, C. L., & Gretebeck, R. J. (2004). Change in diet, physical activity, and body weight in female college freshman. *American Journal of Health Behavior*, 28, 24-32. doi.org/10.5993/AJHB.28.1.3
- Calabrese, J. R., Keck Jr, P. E., Macfadden, W., Minkwitz, M., Ketter, T. A., Weisler, R. H., & Bolder Study Group. (2005). A randomized, double-blind, placebo-controlled trial of quetiapine in the treatment of bipolar I or II depression. *American Journal of Psychiatry*, 162, 1351-1360. doi.org/10.1176/appi.ajp.162.7.1351
- Campbell, D. T. & Kenny, D. A. (1999). *A primer on regression artifacts*. New York, NY: Guilford Press.
- Campbell, D. T., & Boruch, R. F. (1975). Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In: C. A. Bennett., & A. A. Lumsdaine. (Eds.) *Evaluation and experiment: Some critical issues in assessing social programs*, (pp. 195-296). New York, NY: Academic Press.
- Centers for Disease Control and Prevention. (2017). What are the risk factors for lung cancer? Retrieved from https://www.cdc.gov/cancer/lung/basic_info/risk_factors.htm

- Cicirelli, V. G. (1969). *The impact of head start: An evaluation of the effects of head start on children's cognitive and affective development*. (Report presented to the Office of Economic Opportunity Pursuant to Contract of B89-4536, Vols. 1 and 2). Athens, OH: Westinghouse Learning Corporation.
- Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, 27, 724–750. doi.org/10.1002/pam.20375
- Cuijpers, P., Van Straten, A., & Smit, F. (2006). Psychological treatment of late-life depression: A meta-analysis of randomized controlled trials. *International Journal of Geriatric Psychiatry*, 21, 1139-1149. doi: 10.1002/gps.1620
- D'Alonzo, K. T. (2004). The Johnson-Neyman procedure as an alternative to ANCOVA. *Western Journal of Nursing Research*, 26, 804-812. doi.org/10.1177/0193945904266733
- Deptula, D. P., Henry, D. B., & Schoeny, M. E. (2010). How can parents make a difference? Longitudinal associations with adolescent sexual behavior. *Journal of Family Psychology*, 24, 731. doi:10.1037/a0021760
- Depue, R. A. (Ed.). (1979). *The psychobiology of the depressive disorders: Implications for the effects of stress* (Vol. 22). Academic Press.
- Diaz, J. J., & Handa, S. (2006). An assessment of propensity score matching as a nonexperimental impact estimator evidence from Mexico's PROGRESA program. *Journal of Human Resources*, 41, 319-345. doi: 10.3368/jhr.XLI.2.319

- Duncan, G. J., Engel, M., Claessens, A., & Dowsett, C. J. (2014). Replication and robustness in developmental research. *Developmental Psychology*, 50, 2417–2425. doi.org/10.1037/a0037996
- Filla, C., Hays, N. P., Gonzales, D., & Hakkak, R. (2013). Self-reported changes in weight, food intake, and physical activity from high school to college. *Journal of Nutritional Disorders & Therapy*, 3, 2161-0509. doi:10.4172/2161-0509.1000129
- Gu, X. S., & Fraser, M. W. (2014). *Propensity score analysis*, (Vol. 12). Thousand Oaks, CA: Sage Inc.
- Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, 20(1), 102-116.
- Hamilton, B. L. (1976). A Monte Carlo test of the robustness of parametric and nonparametric analysis of covariance against unequal regression slopes. *Journal of the American Statistical Association*, 71, 864-869.
doi: 10.1080/01621459.1976.10480960
- Haviland, A., Nagin, D. S., & Rosenbaum, P. R. (2007). Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods*, 12, 247-267. doi: 10.1037/1082-989X.12.3.247
- Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, 12, 1-145. doi.org/10.3386/w0172
- Heckman, J. (1990). Varieties of selection bias. *American Economic Review*, 80, 313-318.

- Heckman, J. (2010). Selection bias and self-selection. In: S. N. Durlauf S. N., & L. E. Blume (Eds.) *Microeconometrics* (pp. 242-266). The New Palgrave Economics Collection. London: Palgrave Macmillan. doi.org/10.1057/9780230280816_29
- Heckman, J., Ichimura, H., Smith, J., & Todd P. (1998). Characterizing selection bias using experimental data. *Econometrica*, 5, 1017-1098. doi.org/10.3386/w6699
- Hoffman, L. (2015). *Longitudinal analysis: Modeling within-person fluctuation and change*. NY: Routledge.
- Holland, P. W., & Rubin, D. B. (1983). On Lord's paradox. In H. Wainer & S. Messick (Eds.), *Principals of Modern Psychological Measurement* (pp. 3–25). Hillsdale, NJ: Erlbaum. doi.org/10.1002/j.2333-8504.1982.tb01321.x
- Hollingsworth, H. H. (1980). An analytical investigation of the effects of heterogeneous regression slopes in analysis of covariance. *Educational and Psychological Measurement*, 40, 611-618. doi.org/10.1177/001316448004000306
- Huitema, B. E. (2011). *The analysis of covariance and alternatives*. New York: Wiley.
- Huynh, H. (1978). Some approximate tests for repeated measurement designs. *Psychometrika*, 43, 161-175. doi.org/10.1007/BF02293860
- Jung, T., & Wickrama, K. A. S. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass*, 2, 302-317. doi:10.1111/j.1751-9004.2007.00054.x
- Kaplan, G. A., Roberts, R. E., Camacho, T. C., & Coyne, J. C. (1987). Psychosocial predictors of depression: Prospective evidence from the human population laboratory studies. *American Journal of Epidemiology*, 125, 206-220. doi.org/10.1093/oxfordjournals.aje.a114521

- Katon, W. J., Von Korff, M., Lin, E. H., Simon, G., Ludman, E., Russo, J., & Bush, T. (2004). The pathways study: A randomized trial of collaborative care in patients with diabetes and depression. *Archives of General Psychiatry*, *61*, 1042-1049. doi:10.1001/archpsyc.61.10.1042
- Keppel, G., & Wiczens, T. D. (2004). *Design and Analysis: A Researcher's Handbook*. Upper Saddle River, N. J.: Pearson Prentice Hall.
- Kraemer, Helena Chmura, Stice, Eric, Kazdin, Alan, Offord, David, & Kupfer, David. (2001). How do risk factors work together? Mediators, moderators, and independent, overlapping, and proxy risk factors. *American Journal of Psychiatry*, *158*, 848-856. doi.org/10.1176/appi.ajp.158.6.848
- Larzelere, R. E., & Cox, R. B. (2013). Making valid causal inferences about corrective actions by parents from longitudinal data. *Journal of Family Theory & Review*, *5*, 282-299. doi.org/10.1111/jftr.12020
- Larzelere, R. E., Ferrer, E., Kuhn, B. R., & Danelia, K. (2010). Differences in causal estimates from longitudinal analyses of residualized versus simple gain scores: Contrasting controls for selection and regression artifacts. *International Journal of Behavioral Development*, *34*, 180-189. doi.org/10.1177/0165025409351386
- Larzelere, R. E., Lin, H., Payton, M. E., & Washburn, I. J. (in press). Longitudinal biases against corrective actions. *Archives of Scientific Psychology*.
- Larzelere, R. E., Washburn, I. J., Lin, H., & Cox, R. B., Jr. (2017). *Trying to overcome selection bias in longitudinal analyses of corrective actions by professionals*. Paper presented at the Society for Research in Child Development, Austin, TX.

- Levy, K. (1980). A Monte Carlo study of analysis of covariance under violations of the assumptions of normality and equal regression slopes. *Educational and Psychological Measurement*, 40, 835-840. doi.org/10.1177/001316448004000404
- Lewis, D. (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.
- Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, 68, 304–305. doi.org/10.1037/h0025105
- Lord, F. M. (1969). Statistical adjustments when comparing preexisting groups. *Psychological Bulletin*, 72, 336–337. doi.org/10.1037/h0028108
- Marsh, H. W. & Hau K. T. (2002). Multilevel modeling of longitudinal growth and change: Substantive effects or regression toward the mean artifacts? *Multivariate Behavioral Research*, 37, 245-282. doi.org/10.1207/s15327906mbr3702_04
- Maris, E. (1998). Covariance adjustment versus gain scores—revisited. *Psychological Methods*, 3, 309-327. doi:10.1037/1082-989X.3.3.309
- Maxwell, S. E. (1980). Pairwise multiple comparisons in repeated measures designs. *Journal of Educational Statistics*, 5, 269-287.
doi.org/argo.library.okstate.edu/10.3102/10769986005003269
- May, K., & Hittner, J. B. (2010). Reliability and validity of gain scores considered graphically. *Perceptual and Motor Skills*, 111, 399-406.
doi.org/10.2466/03.pms.111.5.399-406
- Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology*, 110, 40. doi.org/10.1037/0021-843X.110.1.40

- Morgan, S. & Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge University Press.
doi.org/10.1017/cbo9780511804564
- Müller, N., Schwarz, M. J., Dehning, S., Douhe, A., Cerovecki, A., Goldstein-Müller, B., & Möller, H. J. (2006). The cyclooxygenase-2 inhibitor celecoxib has therapeutic effects in major depression: Results of a double-blind, randomized, placebo controlled, add-on pilot study to reboxetine. *Molecular Psychiatry*, 11, 680.
doi:10.1038/sj.mp.4001805
- Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, 14, 535-569.
doi.org/10.1080/10705510701575396
- Parekh, R. (2017). What is depression? Retrieved from American Psychiatric Association website: <https://www.psychiatry.org/patients-families/depression/what-is-depression>
- Porter, A. C., & Raudenbush, S. W. (1987). Analysis of covariance: Its model and use in psychological research. *Journal of Counseling Psychology*, 34, 383–392.
doi.org/10.1037/0022-0167.34.4.383
- Rausch, J. R., Maxwell, S. E., & Kelley, K. (2003). Analytic methods for questions pertaining to a randomized pretest, posttest, follow-up design. *Journal of Clinical Child and Adolescent Psychology*, 32, 467–486.
doi.org/10.1207/s15374424jccp3203_15

- Reeves, B. C., Deeks, J. J., Higgins, J., & Wells, G. A. (2008). Including non-randomized studies. In J. P. T. Higgins & S. Green (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions: Cochrane Book Series*, (pp. 389-432). Atrium, England: Cochrane Collaboration and John Wiley & Sons Ltd.
- Reichardt, C. S. (1979). The statistical analysis of data from nonequivalent group designs. In T. D. Cook & D. T. Campbell (Eds.), *Quasi-experimentation: Design and analysis issues for field settings* (pp. 147–205). Boston, MA: Houghton Mifflin.
- Rogosa, D. (1980). Comparing nonparallel regression lines. *Psychological Bulletin*, 88, 307–321. doi.org/10.1037/0033-2909.88.2.307
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.
doi.org/10.1017/cbo9780511810725.016
- Rosenbaum, P. R., & Rubin, D. B. (1984). Estimating the effects caused by treatments: Discussion of a paper by Pratt and Schlaiffer. *Journal of the American Statistical Association*, 79, 26-28.
- Rubin, D. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66, pp. 688–701.
doi.org/10.1037/h0037350
- Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2, 1–26. doi.org/10.1017/cbo9780511810725.009
- Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6, 34-58. doi.org/10.1214/aos/1176344064

- Rubin, D. B. (2004). Teaching statistical inference for causal effects in experiments and observational studies. *Journal of Educational and Behavioral Statistics*, 29, 343–367. doi.org/10.3102/10769986029003343
- Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100, 322–331. doi.org/10.1198/016214504000001880
- Rubin, D. B. (2006). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, 26, 20–36. doi.org/10.1002/sim.2739
- Shadish, W. R. (2010). Campbell and Rubin: A primer and comparison of their approaches to causal inference in field settings. *Psychological methods*, 15, 3-17. doi:10.1037/a0015916
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.
- Selig, J. P., & Little, T. D. (2012). Autoregressive and cross-lagged panel analysis for longitudinal data. In B. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 265-278). New York: Guilford Press.
- Senn, S. J. (2006). Change from baseline and analysis of covariance revisited. *Statistics in Medicine*, 25, 4334–4344. doi.org/10.1002/sim.2682
- Shingles, R. D. (1985). Causal inference in cross-lagged panel analysis. In H. M. Blalock Jr. (Ed.), *Causal models in panel and experimental designs*. New York: Aldine.

- Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological methods*, 15, 250-267. doi:10.1037/a0018719
- Sullivan, L. M., & D'Agostino, R. B., Sr. (2002). Robustness and power of analysis of covariance applied to data distorted from normality by floor effects: Non-homogeneous regression slopes. *Journal of Statistical Computation and Simulation*, 72, 141-165.
- Tripepi, G., Jager, K. J., Dekker, F. W., & Zoccali, C. (2010). Selection bias and information bias in clinical research. *Nephron Clinical Practice*, 115, 94-99. doi.org/10.1159/000312871
- Van Breukelen, G. J. P. (2006). ANCOVA versus change from baseline: More power in randomized studies, more bias in nonrandomized studies. *Journal of Clinical Epidemiology*, 59, 920–925. DOI: 10.1016/j.jclinepi.2006.02.007
- Van Breukelen G. J. P. (2013). ANCOVA versus change from Baseline in Nonrandomized Studies: The Difference, *Multivariate Behavioral Research*, 48, 895-922, doi:10.1080/00273171.2013.831743
- Wainer, H. (1991). Adjusting for differential base rates: Lord's paradox again. *Psychological Bulletin*, 109, 147–151.
- Wainer, H., & Brown, L. (2007). Three statistical paradoxes in the interpretation of group differences: Illustrated with medical school admission and licensing data. *Handbook of Statistics*, 26, 893–918. doi.org/10.1016/S0169-7161(06)26028-0
- Weisberg, H. I. (1979). Statistical adjustments and uncontrolled studies. *Psychological Bulletin*, 86, 1149–1164. doi.org/10.1037/0033-2909.86.5.1149

- Wright, D. B. (2006). Comparing groups in a before-after design: When *t* test and ANCOVA produce different results. *British Journal of Educational Psychology*, 76, 663–675. doi.org/10.1348/000709905X52210
- Wu, P., & Campbell, D. T. (1996). Extending latent variable Lisrel analyses of the 1969 westinghouse head start evaluation to blacks and full year whites. *Evaluation & Program Planning*, 19(3), 183-191. doi.org/10.1016/0149-7189(96)00010-9

Table 1

Research Question Summary

Research Questions	Test in simulation	Test in FFCW
Consistency in Two wave model	Lord's paradox and reversed	Wave 4 and Wave 5 data
Consistency in three-wave model	No	Wave 3, 4, and 5 data
Assumption Tests		
1. Normality	Simulating inflated distribution	Distribution Figure, Hurdle model
2. Homogeneous slopes	Simulating equal vs. unequal	Interaction effect
3. Homogeneous <i>SDs</i>	Simulating equal vs. unequal	No
4. Homogeneous covariates	Simulating equal vs. unequal	<i>t</i> -test
Solution comparison		
5. Latent Class Growth Model (LCGM)		
6. LCGM + group-centered ANCOVA (C-ANCOVA)		
7. LCGM +Propensity Score Methods (PSM)		
8. LCGM + CANCOVA+PSM		

Table 2

Results Simulating Lord's Original and Reversed Paradox Varying Distribution, Slopes, and Standard Deviations

Data Specifications ^a					Means (Pre & Post) ^b				Results ^c	
Normality	SD(f)	SD(m)	Corr(f)	Corr(m)	y0(f)	y0(m)	y1(f)	y1(m)	d	b ₁
Assuming the simple gain score null H ₀ is correct										
Yes	15	15	0.48	0.48	130.02	160.00	129.99	159.99	-0.02	-15.60***
Yes	15	15	0	0.48	130.02	160.00	129.98	159.99	-0.03	-22.81***
Yes	15	15	0	0.96	130.02	160.00	129.98	160.00	-0.04	-15.62***
Yes	15	15	0.48	0	130.02	160.01	129.99	159.99	-0.01	-22.79***
Yes	15	15	0.96	0	130.01	160.01	130.00	159.99	0.01	-15.58***
Yes	5	15	0.48	0.48	130.01	160.00	130.00	159.99	0.00	-15.60***
Yes	5	15	0	0.48	130.01	160.00	129.99	159.99	0.00	-17.05***
Yes	5	15	0	0.96	130.01	160.00	129.99	160.00	-0.01	-4.10***
Yes	5	15	0.48	0	130.01	160.01	130.00	159.99	0.01	-28.54***
Yes	5	15	0.96	0	130.00	160.01	130.00	159.99	0.01	-27.09***
No	15	15	0.48	0.48	85.00	92.50	84.99	92.48	0.00	-0.19
No	15	15	0	0.48	85.00	92.50	85.01	92.48	0.01	-0.19
No	15	15	0	0.96	85.00	92.49	84.99	92.49	-0.01	-0.20
No	15	15	0.48	0	85.00	92.50	85.00	92.48	0.01	-0.19
No	15	15	0.96	0	85.00	92.50	85.00	92.48	0.01	-0.18
No	5	15	0.48	0.48	85.00	92.50	85.00	92.48	0.01	-0.10
No	5	15	0	0.48	85.00	92.50	85.00	92.48	0.01	-0.09
No	5	15	0	0.96	85.00	92.49	85.00	92.49	-0.00	-0.03
No	5	15	0.48	0	85.00	92.50	82.00	92.48	0.01	-0.19
No	5	15	0.96	0	85.00	92.50	85.00	92.48	0.02	-0.18
Assuming the ANCOVA null H ₀ is correct										
Yes	15	15	0.48	0.48	129.99	160.01	137.80	152.20	15.61***	0.02
Yes	5	15	0.48	0.48	130.00	160.01	137.80	152.20	15.61***	0.05
No	15	15	0.48	0.48	84.99	92.50	86.94	90.54	3.91***	3.57***
No	5	15	0.48	0.48	85.00	92.50	86.94	90.54	3.91***	3.65***

Data Specifications ^a					Means (Pre & Post) ^b				Results ^c	
Normality	<i>SD</i> (f)	<i>SD</i> (m)	<i>Corr</i> (f)	<i>Corr</i> (m)	y0(f)	y0(m)	y1(f)	y1(m)	<i>d</i>	<i>b</i> ₁
Equal pretest means; assuming alternative H _A is correct										
Yes	15	15	0.48	0.48	145.02	145.00	129.99	159.99	-30.02***	-30.01***
Yes	15	15	0	0.48	145.02	145.00	129.98	159.99	-30.03***	-30.02***
Yes	15	15	0	0.96	145.02	145.00	129.98	160.00	-30.04***	-30.03***
Yes	15	15	0.48	0	145.02	145.01	129.99	159.99	-30.01***	-30.01***
Yes	15	15	0.96	0	145.01	145.01	130.00	159.99	-30.00***	-29.99***
Yes	5	15	0.48	0.48	145.01	145.00	130.00	159.99	-30.00***	-30.00***
Yes	5	15	0	0.48	145.01	145.00	129.99	159.99	-30.00***	-30.00***
Yes	5	15	0	0.96	145.01	145.00	129.99	160.00	-30.01***	-30.01***
Yes	5	15	0.48	0	145.01	145.01	130.00	159.99	-29.99***	-30.00***
Yes	5	15	0.96	0	145.00	145.01	130.00	159.99	-29.99***	-29.99***
No	15	15	0.48	0.48	88.75	88.75	85.00	92.48	-7.49***	-7.49***
No	15	15	0	0.48	88.75	88.75	85.00	92.48	-7.50***	-7.49***
No	15	15	0	0.96	88.75	88.74	84.99	92.49	-7.51***	-7.50***
No	15	15	0.48	0	88.75	88.75	82.00	92.49	-5.99*	-7.49***
No	15	15	0.96	0	88.75	88.75	85.00	92.48	-7.49***	-7.49***
No	5	15	0.48	0.48	88.75	88.75	85.00	92.48	-7.49***	-7.49***
No	5	15	0	0.48	88.75	88.75	85.00	92.48	-7.50***	-7.49***
No	5	15	0	0.96	88.75	88.74	85.00	92.49	-7.50***	-7.50***
No	5	15	0.48	0	88.75	88.75	85.00	92.48	-7.50***	-7.49***
No	5	15	0.96	0	88.75	88.75	85.00	92.48	-7.48***	-7.49***
Equal pretest means; assuming the null H ₀ is correct (for either simple gain scores or ANCOVA)										
Yes	15	15	0.48	0.48	144.99	145.01	145.00	145.00	0.01	0.00
Yes	5	15	0.48	0.48	145.00	145.01	145.00	145.00	0.01	0.00
No	15	15	0.48	0.48	88.74	88.75	88.74	88.74	0.01	0.00
No	5	15	0.48	0.48	88.75	88.75	88.74	88.74	0.01	0.01

Note. (m) = males, (f) = females. The first boldface line is Lord's paradox and the second boldface line is Lord's paradox reversed.

^aNormality = normal distribution, *SD* = standardized deviation, *Corr* = correlation between pretest and posttest.

^by0 = mean pretest weight, y1 = mean posttest weight. ^c*d* = group difference in simple gain score, *b*₁ = group difference according to ANCOVA.

p* < .05. **p* < .001.

Table 3

Simulation Results on Varying Pretest and Posttest Different for Females

Data setting		Simulated data				Results	
y0(f)	y1(m)	y0(f)	y0(m)	y1(f)	y1(m)	d	b_1
130	130	130.02	160.00	129.99	159.99	-0.02	-15.60***
135	130	135.02	160.00	129.99	159.99	-5.02***	-18.00***
140	130	140.02	160.00	129.99	159.99	-10.02***	-20.40***
145	130	145.02	160.00	129.99	159.99	-15.02***	-22.81***
150	130	150.02	160.00	129.99	159.99	-20.02***	-25.21***
155	130	155.02	160.00	129.99	159.99	-25.02***	-27.61***
160	130	160.02	160.00	129.99	159.99	-30.02***	-30.01***
<i>130</i>	<i>135</i>	<i>130.02</i>	<i>160.00</i>	<i>134.99</i>	<i>159.99</i>	<i>4.98***</i>	<i>-10.60***</i>
135	135	135.02	160.00	134.99	159.99	-0.02	-13.00***
140	135	140.02	160.00	134.99	159.99	-5.02***	-15.40***
145	135	145.02	160.00	134.99	159.99	-10.02***	-17.81***
150	135	150.02	160.00	134.99	159.99	-15.02***	-20.21***
155	135	155.02	160.00	134.99	159.99	-20.02***	-22.61***
160	135	160.02	160.00	134.99	159.99	-25.02***	-25.01***
<i>130</i>	<i>140</i>	<i>130.02</i>	<i>160.00</i>	<i>139.99</i>	<i>159.99</i>	<i>9.98***</i>	<i>-5.60***</i>
<i>135</i>	<i>140</i>	<i>135.02</i>	<i>160.00</i>	<i>139.99</i>	<i>159.99</i>	<i>4.98***</i>	<i>-8.00***</i>
140	140	140.02	160.00	139.99	159.99	-0.02	-10.40***
145	140	145.02	160.00	139.99	159.99	-5.02***	-12.81***
150	140	150.02	160.00	139.99	159.99	-10.02***	-15.21***
155	140	155.02	160.00	139.99	159.99	-15.02***	-17.61***
160	140	160.02	160.00	139.99	159.99	-20.02***	-20.01***
<i>130</i>	<i>145</i>	<i>130.02</i>	<i>160.00</i>	<i>144.99</i>	<i>159.99</i>	<i>14.98***</i>	<i>-0.60</i>
<i>135</i>	<i>145</i>	<i>135.02</i>	<i>160.00</i>	<i>144.99</i>	<i>159.99</i>	<i>9.98***</i>	<i>-3.00*</i>
<i>140</i>	<i>145</i>	<i>140.02</i>	<i>160.00</i>	<i>144.99</i>	<i>159.99</i>	<i>4.98***</i>	<i>-5.40***</i>
145	145	145.02	160.00	144.99	159.99	-0.02	-7.81***
150	145	150.02	160.00	144.99	159.99	-5.02***	-10.21***
155	145	155.02	160.00	144.99	159.99	-10.02***	-12.61***
160	145	160.02	160.00	144.99	159.99	-15.02***	-15.01***
130	150	130.02	160.00	149.99	159.99	19.98***	4.40**

Data setting		Simulated data				Results	
y0(f)	y1(m)	y0(f)	y0(m)	y1(f)	y1(m)	<i>d</i>	<i>b</i> ₁
135	150	135.02	160.00	149.99	159.99	14.98***	2.00
<i>140</i>	<i>150</i>	<i>140.02</i>	<i>160.00</i>	<i>149.99</i>	<i>159.99</i>	<i>9.98***</i>	<i>-0.40</i>
<i>145</i>	<i>150</i>	<i>145.02</i>	<i>160.00</i>	<i>149.99</i>	<i>159.99</i>	<i>4.98***</i>	<i>-2.81*</i>
150	150	150.02	160.00	149.99	159.99	-0.02	-5.21***
155	150	155.02	160.00	149.99	159.99	-5.02***	-7.61***
160	150	160.02	160.00	149.99	159.99	-10.02***	-10.01***
130	155	130.02	160.00	154.99	159.99	24.98***	9.40***
135	155	135.02	160.00	154.99	159.99	19.98***	7.00***
140	155	140.02	160.00	154.99	159.99	14.98***	4.60***
145	155	145.02	160.00	154.99	159.99	9.98***	2.19 ^a
<i>150</i>	<i>155</i>	<i>150.02</i>	<i>160.00</i>	<i>154.99</i>	<i>159.99</i>	<i>4.98***</i>	<i>-0.21</i>
155	155	155.02	160.00	154.99	159.99	-0.02	-2.61*
160	155	160.02	160.00	154.99	159.99	-5.02***	-5.01***
130	160	130.02	160.00	159.99	159.99	29.98***	14.40***
135	160	135.02	160.00	159.99	159.99	24.98***	12.00***
140	160	140.02	160.00	159.99	159.99	19.98***	9.60***
145	160	145.02	160.00	159.99	159.99	14.98***	7.19***
150	160	150.02	160.00	159.99	159.99	9.98***	4.79***
155	160	155.02	160.00	159.99	159.99	4.98***	2.39*
160	160	160.02	160.00	159.99	159.99	-0.02	-0.01

Note. Data setting: males pretest/posttest mean = 160, slope for males/females = 0.48, standard deviation = 15.

The italic font are results in opposite directions.

The first bold font is Lord's paradox and the second bold font is Lord's paradox reversed.

(m) = males; (f) = females; y0 = pretest weight; y1 = posttest weight; *d* = CHANGE approach; *b*₁ = ANCOVA approach.

^a*p* < .10, **p* < .05, ***p* < .01, ****p* < .001

Table 4

Simulation Results on Paradox and Reversed Paradox for Comparing Difference Approaches

data	Simulated data				Results			
	y0(f)	y0(m)	y1(f)	y1(m)	d	$t(d)$	b_1	$t(b_1)$
Lord 's paradox	130.02	160.00	129.99	159.99				
Original scale					-0.02	-0.002	-15.60***	15.50
Centered scale					-0.01	-0.01	- 0.01	-0.01
Reversed paradox	129.99	160.01	137.80	152.20				
Original scale					15.61***	16.17	0.02	0.02
Centered scale					15.61***	16.17	15.61***	18.76

Note. The original scale is the original simulated pretest scores and posttest scores. For the centered scale, both the original pretest and posttest scores are centered around the *pretest* group mean. (m) = males; (f) = females; y0 = pretest weight; y1 = posttest weight; d = treatment effect using the simple gain score approach; b_1 = treatment effect using the ANCOVA approach; $t(d)$ = the t -test score for d ; $t(b_1)$ = the t -test score for b_1 .

*** $p < .001$.

Table 5

Simulated Data for Comparing Results from Original and Matched Data within Sub-Groups

	d_0	d_1	b_1	N_f/N_m
One selected data				
Lord's paradox	-30.42***	- 0.02	-15.99***	500/50
≥ 145 pound group				
Original data	-13.02***	-7.26***	-13.67***	423/87
Matched sample	-1.51 ^a	-11.77***	-12.74***	77/154
< 145 pound group				
Original data	-11.21***	-12.37***	-18.37***	77/414
Matched sample	-0.77	-18.75***	-19.11***	172/86
Reversed Lord's paradox	-30.32***	15.60***	0.03	500/500
≥ 145 pound group				
Original data	-11.60***	7.37***	0.72	81/430
Matched sample	-0.68	1.41	0.91	81/162
< 145 pound group				
Original data	-10.29***	3.41 ^a	-1.05	419/70
Matched sample	-0.03	-2.17	-2.15	140/70
Mean of 1000 data				
Lord's paradox	-29.99***	-0.02	-15.61***	500/500
≥ 145 pound group				
Original data	-11.45***	-9.60***	-15.55***	423/87
Matched sample	-1.06	-15.00***	-15.55***	77/154
< 145 pound group				
Original data	-11.38***	-9.72***	-15.64***	77/414
Matched sample	-1.00	-15.10***	-15.63***	172/86
Reversed Lord's paradox	-30.02***	15.61***	0.02	500/500
≥ 145 pound group				0
Original data	-11.44***	6.00*	0.07	81/430
Matched sample	-1.07	0.64	0.09	81/162
< 145 pound group				
Original data	-11.46***	6.00*	0.01	419/70
Matched sample	-1.05	0.57	0.02	140/70

Note. The original data is the original simulated pretest scores and posttest scores. The Matched sample is the sample with males and females matched on the pretest condition after splitting the original data into two parts.

N_f = females' sample size; N_m = males' sample size; d_0 = pretest weight difference between males and females; d_1 = treatment effect using simple gain score approach; b_1 = treatment effect using ANCOVA approach.

* $p < .05$. *** $p < .001$.

Table 6

Frequency of Mothers' depression severity score

MD Score	severity score	Wave 2	Wave 3	Wave 4	Wave 5
0	0	3233	2921	3056	2596
	1	183	173	132	127
	2	165	177	157	125
	3	26	34	36	21
	4	23	26	36	18
1	5	40	8	7	6
2	6	29	18	20	27
3	7	62	48	45	60
4	8	102	107	80	54
5	9	167	192	156	123
6	10	169	248	185	162
7	11	114	198	170	148
8	12	51	71	59	48
Totals		4364	4221	4139	3515

Note. MD score = Major Depression score, i.e., number of diagnostic symptoms endorsed after stem questions answered to qualify for possible Major Depression.

Table 7

Adjusted Effects and Correlations between Covariates, Treatments, and Outcomes

Covariates	Psychological Treatment					Medication Treatment			
	$r(z_{y1})^b$	$r(z_{x_{psy}})^c$	$r(z_{y1}) \times r(z_{x_{psy}})$	Adjusted ^d	Relative ^e %	$r(z_{x_{med}})^f$	$r(z_{y1}) \times r(z_{x_{med}})$	Adjusted ^d	Relative ^e %
<i>Depress 4</i>	0.35***	0.38***	0.1330	1.74	50.04	0.32***	0.1600	1.79	47.75
<i>Depress 3</i>	0.33***	0.20***	0.0660	2.65	24.00	0.21***	0.0420	2.37	30.99
<i>Depress 2</i>	0.24***	0.14***	0.0336	3.03	13.13	0.17***	0.0408	2.77	19.32
<i>Smoke 2</i>	0.16***	0.10***	0.0160	3.25	6.99	0.10***	0.0160	3.18	7.28
<i>Financial hardship</i>	0.21***	0.14***	0.0294	3.34	4.08	0.11***	0.0231	3.13	8.85
<i>Domestic violence</i>	0.12***	0.08***	0.0096	3.36	3.85	0.05*	0.0060	3.33	2.90
<i>Externalizing</i>	0.15***	0.07***	0.0105	3.36	3.77	0.06*	0.0090	3.04	11.40
<i>Parental stress</i>	0.13***	0.06***	0.0078	3.37	3.38	0.06***	0.0078	3.28	4.41
Alcohol used	0.03*	0.01	0.0003	3.59	2.77	0.03 ^a	0.0009	3.38	1.47
Alcohol (dummy)	0.04*	0.03 ^a	0.0012	3.55	1.70	0.08 ^a	0.0032	3.40	0.91
Drug used	0.10***	0.05**	0.0050	3.56	1.89	0.09***	0.0090	3.38	1.48
<i>Drug (dummy)</i>	0.12***	0.05*	0.0060	3.48	0.23	0.07***	0.0084	3.37	1.82
<i>Foreign born</i>	0.06**	0.07***	0.0042	3.45	1.15	0.08***	0.0048	3.38	1.36
<i>Child health</i>	-0.06*	-0.03*	0.0018	3.46	0.93	-0.05**	0.0030	3.37	1.76
Black	0.03 ^a	-0.03 ^a	0.0009	3.48	0.40	-0.05**	0.0015	3.39	1.26
Hispanic	-0.01	-0.02	0.0002	3.46	0.89	-0.07***	0.0007	3.35	2.22
Other race	-0.04*	0.00	0.0000	3.46	0.91	-0.02	0.0008	3.35	2.45
<i>Social support</i>	-0.13***	-0.05***	0.0065	3.46	0.78	-0.05***	0.0065	3.42	0.39
<i>Partner support</i>	0.14***	0.12***	0.0168	3.51	0.64	0.07***	0.0028	3.32	3.24
<i>Cohabitation</i>	-0.08***	-0.04**	0.0032	3.51	0.53	-0.02	0.0016	3.38	1.41
<i>Mother age</i>	-0.05*	0.03 ^a	0.0015	3.50	0.34	0.05*	0.0025	3.48	1.51
<i>Treatment Hardship</i>	0.10***	0.05*	0.0050	3.47	0.37	0.05*	0.0050	3.39	1.28
<i>Mother health</i>	0.01	0.04*	0.0004	3.80	0.05	0.05*	0.0005	3.60	0.03
<i>(Mother health)²</i>	0.06***	0.07***	0.0042	3.74	1.39	0.10***	0.0060	3.52	2.00
Religion attendance	-0.03	-0.01	0.0003	3.56	1.29	-0.04*	0.0012	3.45	0.69
Child gender	0.02	0.02	0.0004	3.48	0.25	-0.01	0.0050	3.43	0.07
Number of kids	0.03 ^a	0.01	0.0003	3.53	1.01	0.01	0.0003	3.46	0.78
Poverty ratio 1	-0.10***	-0.01	0.0010	3.48	0.21	0.02	-0.0020	3.47	1.13
Poverty ratio 3	-0.09***	-0.01	0.0009	3.59	0.00	0.00	0.0018	3.59	0.00

	Psychological Treatment					Medication Treatment			
Covariates	r(z _{y1}) ^b	r(z _{x_{psy}}) ^c	r(z _{y1})×r(z _{x_{psy}})	Adjusted ^d	Relative ^e %	r(z _{x_{med}}) ^f	r(z _{y1})×r(z _{x_{med}})	Adjusted ^d	Relative ^e %
(Poverty ratio 3) ²	-0.04**	0.02	0.0008	3.78	0.45	0.05**	0.0020	3.63	1.10
Mother education	-0.12***	0.00	0.0000	3.81	0.43	0.01	-0.0012	3.48	1.49

Note: Unadjusted psychological treatment effect is 3.49, and unadjusted medication treatment effect is 3.43. Italicized covariates were included in propensity score calculation.

^b $r(z y_1)$ = Correlation between covariate (z) and depression outcome at wave 5 (y_1);

^c $r(z x_{psy})$ = Correlation between covariate (z) and psychological treatment (x_{psy});

^d The adjusted treatment effect after controlling for the covariate;

^e Relative % = The difference between the unadjusted effect and the adjusted effect, expressed as the % reduction;

^f $r(z x_{med})$ = Correlation between covariate (z) and medication treatment (x_{med}).

* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 8

Effect Sizes for Psychological and Medication Treatments for Mother's Depression

Data	Difference-in-difference		Regression	
	<i>d</i>	<i>S.E.</i>	<i>b</i> ₁	<i>S.E.</i>
Psychological Treatment				
Original scale	−2.307***	0.300	1.744***	0.265
Centered scale	−2.307***	0.300	−2.303***	0.245
Medication Treatment				
Original scale	−1.869***	0.324	1.793***	0.279
Centered scale	−1.866***	0.324	−1.953***	0.264

Note. The original scale is the original pretest scores and posttest scores. The centered scale is centering the original pretest and posttest scores around the pretest group means.

d = Treatment effect using simple gain score approach; *b*₁ = Treatment effect using ANCOVA approach.

****p* < .001.

Table 9

Linear Growth model using a zero-inflated Hurdle Model to Estimate Trajectory Groups

Fit Statistics	1 class	2 class	3 class	4 class
Trajectory Model				
Log Likelihood		-15708.002	-15414.880	-15292.450
AIC	N/A	31400.003	30857.759	30620.900
BIC	N/A	31470.759	30947.811	30736.681
SSA-BIC	N/A	31435.805	30903.325	30679.484
Entropy	N/A	0.703	0.704	0.740
LMR test	N/A	1590.976	446.318	36.761
LMR <i>p</i> -value		< 0.001	< 0.001	< 0.001
Trajectory groups				
<i>Two –class model</i>	<i>1</i>	<i>2</i>		
1, n = 3053, 66.47%	0.904	0.096		
2, n = 1540, 33.53%	0.011	0.989		
<i>Three-class model</i>	<i>1</i>	<i>2</i>	<i>3</i>	
1, n = 3026, 65.88%	0.886	0.099	0.016	
2, n = 1300, 28.30%	0.025	0.934	0.041	
3, n = 267, 5.81%	0.098	0.110	0.793	
<i>Four-class model</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>
1, n = 2964, 64.53%	0.870	0.102	0.025	0.003
2, n = 1208, 26.30%	0.007	0.938	0.054	0.002
3, n = 371, 8.08%	0.074	0.101	0.811	0.014
4, n = 50, 1.09%	0.052	0.069	0.305	0.573

Note. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion;
SSA-BIC = Sample-Size-Adjusted BIC; LMR = Lo–Mendell–Rubin test;
BLRT = Bootstrap Likelihood Ratio Test.
N = 4593.

Table 10

Three Trajectory Groups Estimates Based on Zero-Inflated Hurdle model

	Low depression	High depression	Medium depression
Trajectory model			
Probability	65.88%	28.30%	5.81%
Mean depress at 2	0.22	4.48	4.41
Mean depress at 3	0.30	5.74	3.61
Mean depress at 4	0.36	4.84	3.64
(0 vs. 1) Intercept	0	-2.40***	-2.97***
(0 vs. 1) Slope	0	0.06	-0.62***
(none 0) Intercept	0	2.10***	1.90***
(none 0) Slope	0	0.08***	-0.62***
Trajectory groups			
PSY treatment	1.88%	12.23%	9.74%
MED treatment	1.42%	10.00%	7.87%
Zero depression at 2	91.1%	43.6%	37.40%
Zero depression at 3	90.4%	35.4%	10.9%
Zero depression at 4	90.9%	47.5%	9.4%
Zero depression at 5	81.4%	56.5%	48.3%
Total	3026	1300	267

Note. $N = 4593$.*** $p < .001$.

Table 11

Comparing Results between the Simple Gain Score Approach and the Residual Gain Score Approach from Original and Matched Data within Three Sub-groups for FFCW Data

	d_0	d_1	b_1	N_T/N_C
Low or None depression group				
<i>Psychological treatment</i>				
Imputed data	2.74***	-0.02	1.63***	57/2969
Centered data	0.00	-0.02	-0.02	
Matched on Pre (all)	0.32	1.79*	1.97***	57/228
Matched on Pre (truncated)	0.01	2.25	2.28**	19/76
Matched on Multi (all)	-0.39	1.47 ^a	1.26	57/57
Matched on Multi (truncated)	0.11	-0.37	-0.14	19/19
<i>Medication treatment</i>				
Imputed data	2.58***	-0.18	1.32**	43/2983
Centered data	0.00	-0.18	-0.18	
Matched on Pre (all)	0.09	1.81*	1.86***	43/172
Matched on Pre (truncated)	0.00	1.02	1.02	13/52
Matched on Multi (all)	0.41	0.81	1.04	43/43
Matched on Multi (truncated)	0.15	0.54	0.91	13/13
High depression group				
<i>Psychological treatment</i>				
Imputed data	4.63***	-1.70***	1.31***	159/1141
Centered data	0.00	-1.70***	-1.70***	
Matched on Pre (all)	0.28	1.03*	1.20**	159/477
Matched on Pre (truncated)	0.27	1.00	1.28*	113/339
Matched on Multi (all)	0.35	1.01*	1.25**	159/477
Matched on Multi (truncated)	0.33	1.04 ^a	1.38**	113/339
<i>Medication treatment</i>				
Imputed data	4.30***	-1.64***	1.12**	130/1170
Centered data	0.00	-1.64***	-1.64***	
Matched on Pre (all)	0.21	1.04*	1.16*	130/520
Matched on Pre (truncated)	0.23	1.36*	1.59**	89/356
Matched on Multi (all)	0.37	0.96	1.18*	130/130
Matched on Multi (truncated)	-0.21	1.38 ^a	1.25 ^a	89/89
Medium depression group				
<i>Psychological treatment</i>				
Original data	3.17**	-1.26	1.53 ^a	26/241
Centered data	0.00	-1.26	-1.26	
Matched on Pre (all)	0.85	0.59	1.40	26/102
Matched on Pre (truncated)	0.94	0.28	1.27	21/84
Matched on Multi (all)	0.96	0.19	1.09	26/26
Matched on Multi (truncated)	0.80	-1.57	-0.75	21/21

	d_0	d_1	b_1	N_T/N_C
<i>Medication treatment</i>				
Imputed data	1.94 ^a	1.59	3.30***	21/246
Centered data	0.00	1.59	1.59 ^a	
Matched on Pre (all)	0.11	2.63	2.73*	21/84
Matched on Pre (truncated)	-0.02	2.37	2.34 ^a	19/76
Matched on Multi (all)	-0.71	2.71	1.97	21/21
Matched on Multi (truncated)	-0.84	3.68 ^a	2.85 ^a	19/19

Note: Pre means sample matching from propensity score calculated based on the pretest outcome (Wave 4 depression). Multi means sample matching from propensity score calculated based on the multiple covariates. (all) means including the zero and none-zero depression samples in the analyses. (truncated) means only including Wave 4 and Wave 5 none zero depression samples in the analyses.

N_T = Sample size in treated group; N_C = Sample size in control group; d_0 = Pretest difference between the treatment group and the comparison group; d_1 = Treatment effect using simple gain score approach; b_1 = Treatment effect using ANCOVA approach.

^a $p < .10$. * $p < .05$. ** $p < .01$. *** $p < .001$.

Table 12

Mean and Propensity of Being in the Treatment Group for the 16 covariates (Depression at Wave 4)

Covariates	Low Depression			High Depression			Medium Depression		
	M	PSY	MED	M	PSY	MED	M	PSY	MED
<i>Propensity score calculated from the pretest outcome ONLY</i>									
Depress 4	0.34	0.34***	0.34***	4.93	0.24***	0.22***	3.64	0.21***	0.13*
<i>Propensity of log score calculated from the 16 Covariates</i>									
Depress 4	0.34	0.38***	0.34**	4.93	0.24***	0.22**	3.64	<0.01	<0.01
Depress 3	0.28	0.02***	0.09***	6.01	0.01	0.04 ^a	3.61	<0.01***	<0.01
Depress 2	0.19	-0.04	<0.01	4.37	-0.04 ^a	<0.01	4.41	< 0 ^b	< 0 ^b
Smoke 2	0.27	0.55**	0.72***	0.48	0.15	0.20 ^a	0.46	<0.01*	<0.01
Financial hardship	1.54	-0.04	-0.21 ^a	2.71	0.04	0.01	2.88	<0.01	<0.01
Domestic violence	0.47	-0.13	-0.72 ^a	0.67	0.16	-0.46 ^a	0.71	< 0 ^b	< 0 ^b
Externalizing	0.57	-0.89 ^a	-0.59	0.72	-0.14	0.14	0.73	< 0 ^b	<0.01
Parental stress	2.14	0.55*	-0.14	2.41	-0.08	0.16	2.50	< 0 ^{ab}	< 0 ^b
Drug used	0.002	2.21 ^a	-12.92	0.02	-0.88	0.32	0.02	<0.01*	<0.01
Mother Foreign born	0.81	1.34*	1.23 ^a	0.87	0.40	1.65**	0.86	<0.01	<0.01
Child health	4.51	0.28	-0.35 ^a	4.41	-0.09	-0.12	4.33	< 0* ^b	< 0** ^b
Social support	2.67	-0.04	-0.02	2.36	0.16	0.20 ^a	2.26	<0.01	< 0 ^b
Partner support	1.36	1.91***	0.58	1.52	0.29	-0.18	1.50	<0.01*	<0.01
Cohabitation	0.56	0.35	0.34	0.45	-0.25	-0.02	0.36	< 0 ^b	<0.01
Mother age	25.46	0.06*	0.03	24.60	0.03*	0.06***	25.18	< 0 ^b	<0.01
Mother health	0.05	0.05	0.13	-0.08	0.13 ^a	-0.02	-0.26	< 0 ^b	<0.01
(Mother health) ²	1.18	0.10*	0.07 ^a	1.93	-0.01	0.04*	2.41	< 0 ^b	< 0 ^b

Note. ^b<0 = Negative value very close to zero. A log score very close to zero means close to 50% of chance that a mother would be likely to be in a treatment group.

^a $p < .10$. * $p < .05$. ** $p < .01$. *** $p < .001$.

Table 13

Covariates Balance Based on Standardized Differences (Cohen's d), Before and After Matching within Three Trajectory Groups Using Samples Including Zeros Depression at Wave 4 and 5

	Low Depression		High Depression		Medium Depression	
	Before	After	Before	After	Before	After
<i>Psychological TXT</i>						
Depress 4	0.86	0.08	1.09	0.10	0.80	0.22
Depress 3	0.26	0.01	0.20	0.08	0.06	0.11
Depress 2	0.07	0.24	0.19	0.02	0.16	0.18
Mother smoke 2	0.37	0.17	0.20	0.08	0.55	0.12
Financial hardship	0.16	0.13	0.18	0.06	0.27	0.23
Domestic violence	0.30	0.25	0.11	0.04	0.13	0
Externalizing	0.02	0.19	0.12	0.03	0.08	0.18
Parental stress	0.27	0.03	0.03	0.01	0.28	0.27
Cm3drug_case	0.16	0.19	0.02	0.07	0.26	0.13
Mother Foreign born	0.36	0.07	0.19	0.09	0.61	NA
Child health	0.10	0.09	0.07	0.02	2.04	0.47
Social support	0.22	0.09	0.03	0.04	0.14	0.29
Partner support	0.57	0.15	0.14	0.04	0.29	0.09
Cohabitation	0.01	0.07	0.08	0.05	0.41	0
Mother age	0.28	0.19	0.15	0.06	0.04	0.28
Mother heath	0.23	0.04	0.20	0.10	0.06	0.34
(Mother heath) ²	0.30	0.02	0.10	0.08	0.28	0.54
<i>Medication TXT</i>						
Depress 4	0.81	0.10	0.99	0.10	0.50	0.16
Depress 3	0.27	0.37	0.17	0.07	0.15	0.08
Depress 2	0.02	0.14	0.01	0.18	0.07	0.23
Mother smoke 2	0.42	0.02	0.30	0.08	0.46	0.05
Financial hardship	0.20	0.25	0.20	0.03	0.23	0.04
Domestic violence	0.23	0.27	0.05	0.16	0.31	0.10
Externalizing	0.03	0.12	0.02	0.04	0.16	0.17
Parental stress	0.14	0.23	0.20	0.07	0.15	0.04
Cm3drug_case	0.06	NA	0.16	0.07	0.33	0
Mother Foreign born	0.36	0.16	0.39	0	0.60	NA
Child health	0.15	0.13	0.13	0.09	0.45	0.06
Social support	0.05	0	0.02	0.09	0.12	0.23
Partner support	0.01	0.10	0.07	0.00	0.04	0
Cohabitation	0.13	0.05	0.03	0.06	0.06	0.10
Mother age	0.16	0.05	0.31	0.07	0.12	0.03
Mother heath	0.24	0.09	0.13	0.03	0.11	0.06
(Mother heath) ²	0.21	0.13	0.23	0.03	0.27	0.09

Note: Covariates that are not adequately balanced are bolded.

TXT = treatment; Before = before matching sample; After = after matching sample; NA = the comparison group only has zero scores for that covariate.

The cut point is .20. A standardized difference greater than .20 is considered not balanced.

Table 14

Covariates Balance Based on Standardized Differences (Cohen's d): Before and After Matching within Three Subgroups Using Samples NOT Including Zeros Depression at Wave 4 and 5

	Low Depression		High Depression		Medium Depression	
	Before	After	Before	After	Before	After
<i>Psychological TXT</i>						
Depress 4	0.79	0.06	0.66	0.13	0.77	0.19
Depress 3	0.01	0.43	0.31	0.03	0.09	0.25
Depress 2	0.08	0.09	0.16	0.08	0.12	0.46
Mother smoke 2	0.27	0.06	0.16	0.05	0.63	0.08
Financial hardship	0.27	0.09	0.07	0.00	0.33	0.23
Domestic violence	0.36	0.00	0.02	0.02	0.01	0.35
Externalizing	0.04	0.12	0.03	0.05	0.14	0.04
Parental stress	0.45	0.14	0.05	0.06	0.18	0.12
Cm3drug_case	0.31	0.33	0.09	0.10	0.13	0.00
Mother Foreign born	0.14	0.00	0.29	0.06	0.60	NA
Child health	0.06	0.20	0.02	0.02	0.31	0.15
Social support	0.16	0.10	0.16	0.11	0.17	0.25
Partner support	0.37	0.21	0.05	0.03	0.33	0.31
Cohabitation	0.03	0.21	0.03	0.02	0.42	0.43
Mother age	0.39	0.20	0.06	0.03	0.00	0.22
Mother heath	0.30	0.14	0.16	0.11	0.06	0.14
(Mother heath) ²	0.40	0.20	0.04	0.04	0.31	0.08
<i>Medication TXT</i>						
Depress 4	0.68	0.10	0.75	0.06	0.23	0.19
Depress 3	0.07	0.58	0.19	0.03	0.17	0.13
Depress 2	0.16	0.18	0.14	0.04	0.22	0.01
Mother smoke 2	0.16	0.31	0.23	0.04	0.49	0.12
Financial hardship	0.11	0.27	0.16	0.15	0.22	0.12
Domestic violence	0.12	0.15	0.09	0.15	0.21	0.35
Externalizing	0.06	0.52	0.01	0.13	0.15	0.09
Parental stress	0.39	0.28	0.15	0.08	0.08	0.34
Cm3drug_case	0.09	NA	0.04	0.21	0.38	0.20
Mother Foreign born	0.24	0.24	0.23	0.06	0.60	NA
Child health	0.12	0.09	0.15	0.13	0.52	0.11
Social support	0.28	0.15	0.08	0.01	0.09	0.29
Partner support	0.39	0.28	0.01	0.23	0.11	0.17
Cohabitation	0.10	0.00	0.08	0.02	0.03	0.23
Mother age	0.28	0.00	0.20	0.03	0.21	0.03
Mother heath	0.23	0.04	0.11	0.14	0.01	0.39
(Mother heath) ²	0.39	0.12	0.25	0.11	0.27	0.27

Note: The cut point is .20. A difference score greater than .20 is considered not balanced.

TXT = treatment; Before = before matching sample; After = after matching sample; NA = the comparison group only has zero scores for that covariate.

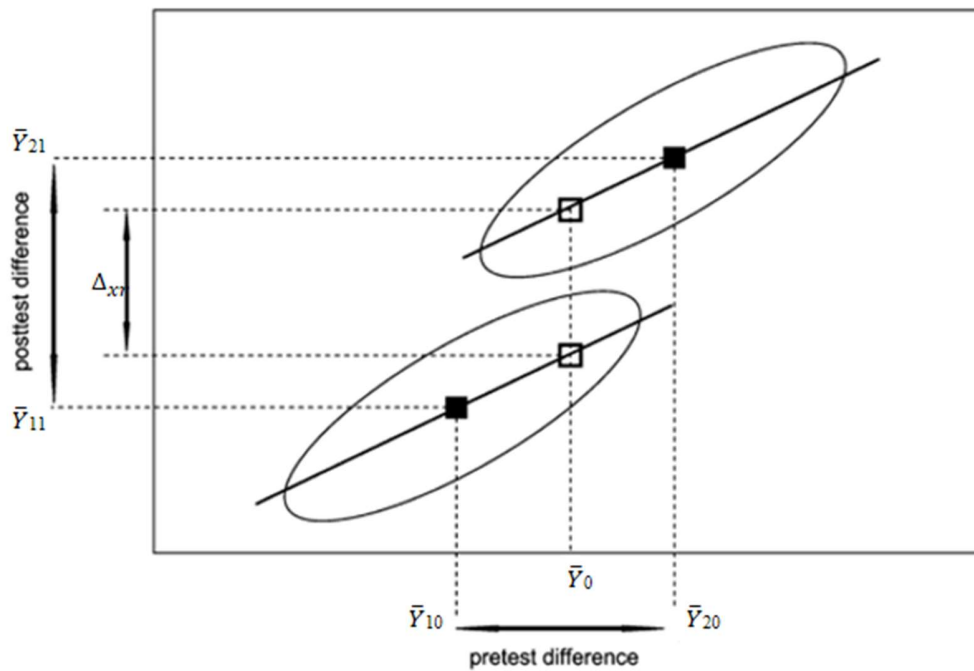


FIGURE 1 ANCOVA adjustment of the posttest group difference for the pretest group difference. Moving from observed means (■) to adjusted means (□) gives the adjusted posttest group difference on the Y -axis.

Figure 1. Van Breukelen's (2013) figure on ANCOVA adjustment of the posttest difference for the pretest difference.

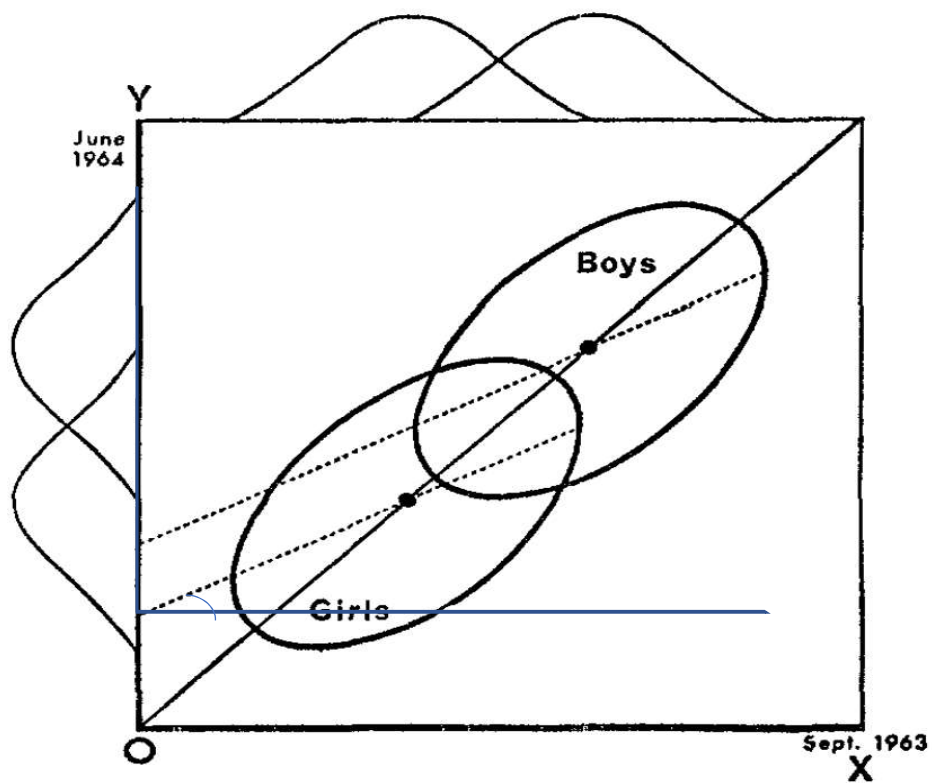


FIG. 1. Hypothetical scatterplots showing initial and final weight for boys and for girls.

Figure 2. Lord's (1967) paradox.

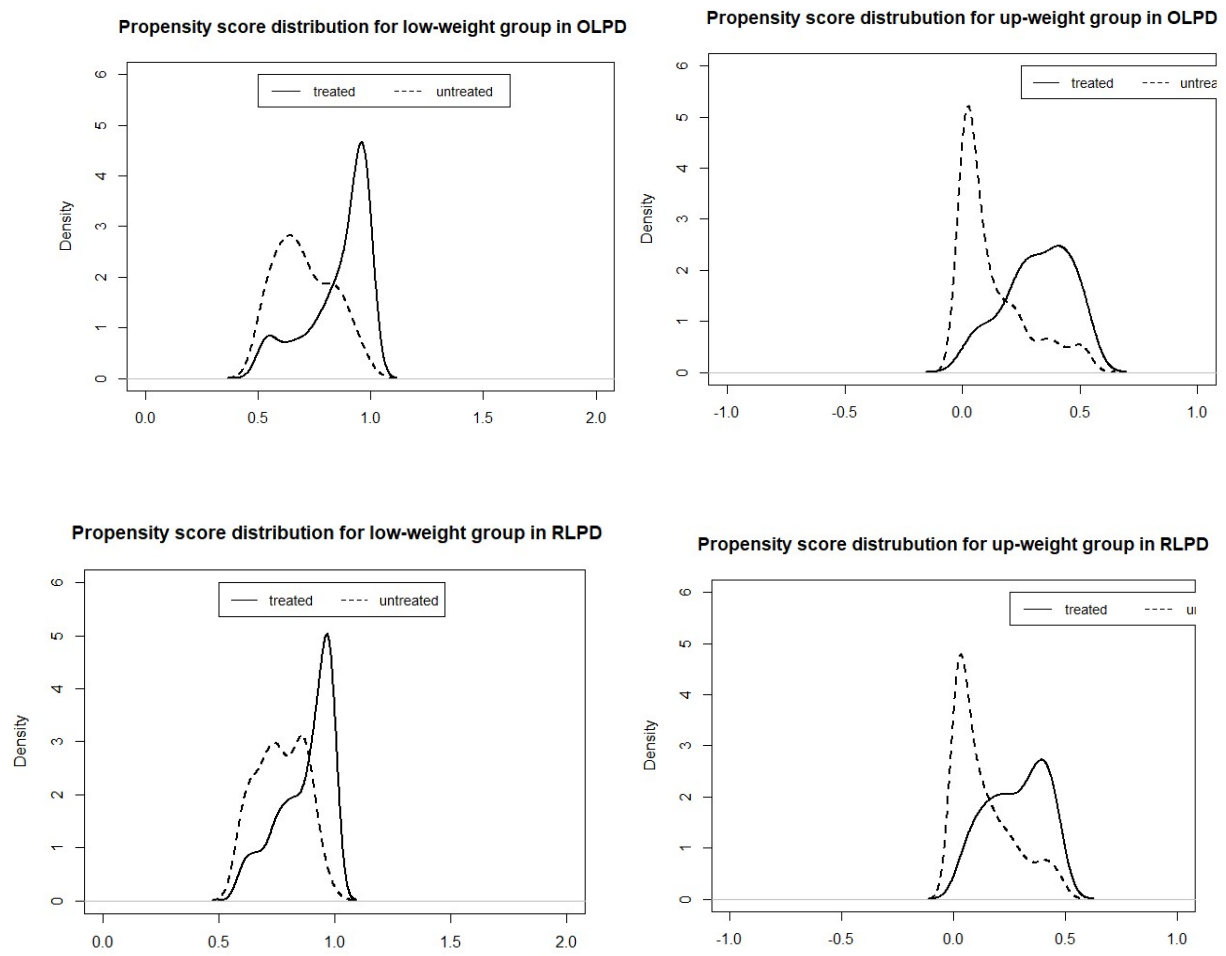
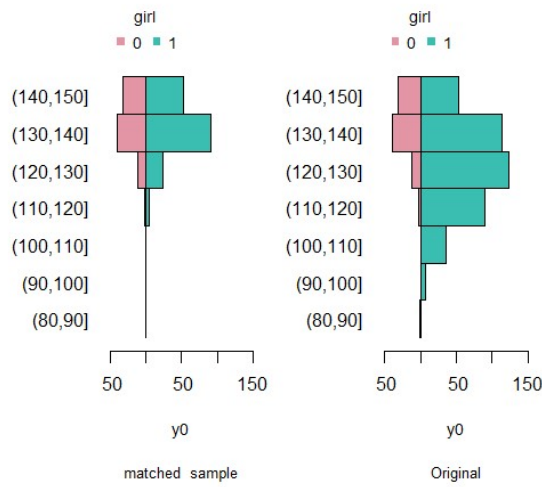
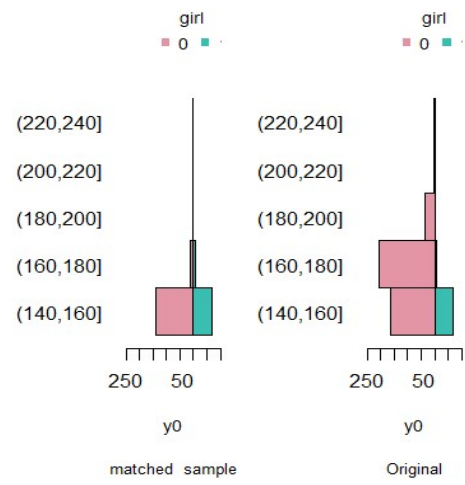


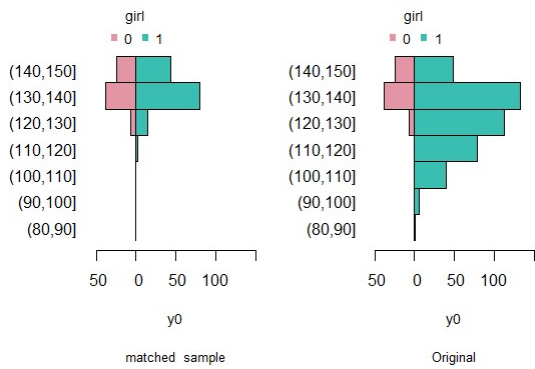
Figure 3. Propensity score distribution for simulated Lord's paradox data (OLPD) and the reversed Lord's paradox data (RLPD).



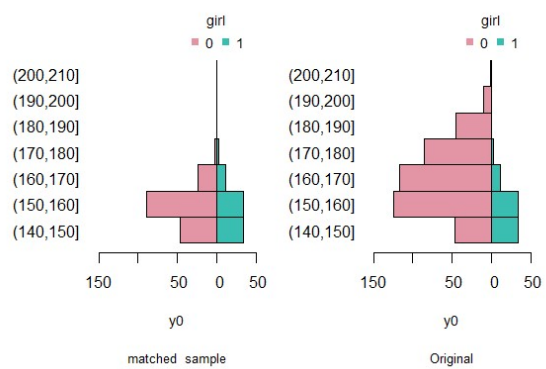
A: low weight sub-group in OLPD



B: high weight sub-group in OLPD



C: low weight sub-group in RLPD



D: high weight sub-group in RLPD

Figure 4. Matched and unmatched samples for simulated Lord's paradox data (OLPD) and the reversed Lord's paradox data (RLPD).

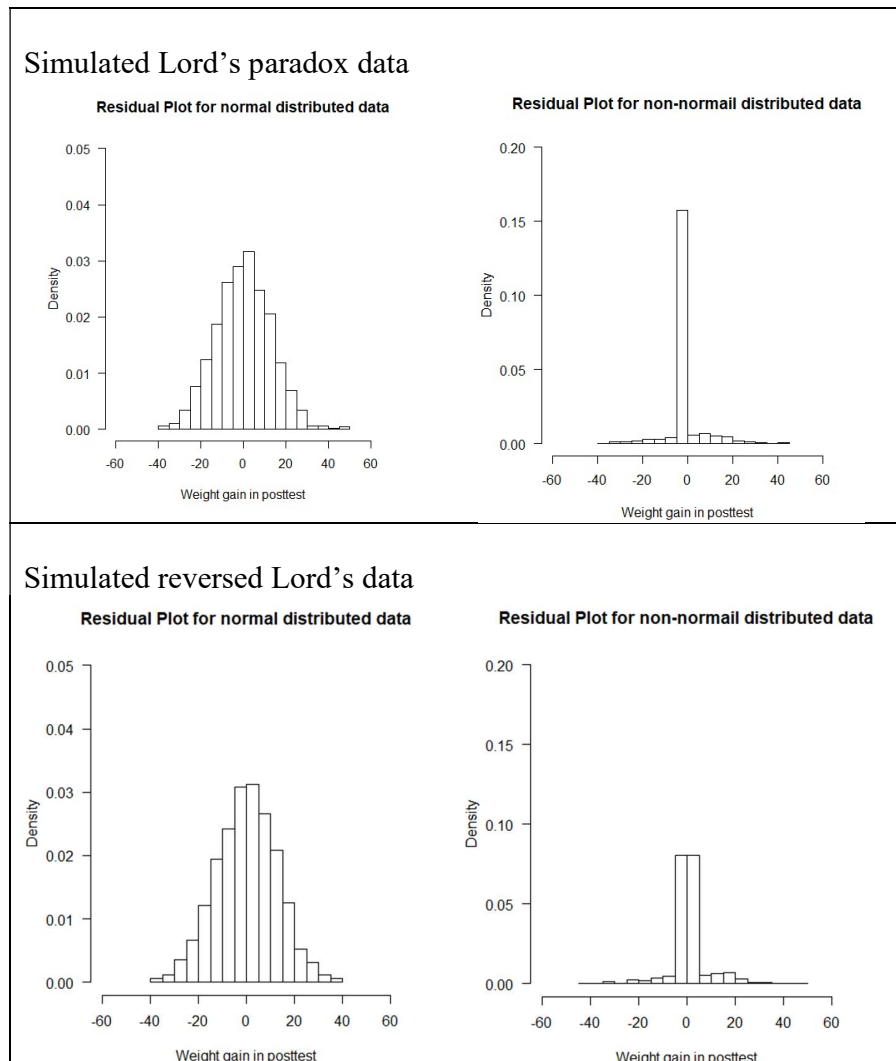


Figure 5. Histogram plots for residuals in normal distributed and non-normal distributed data.

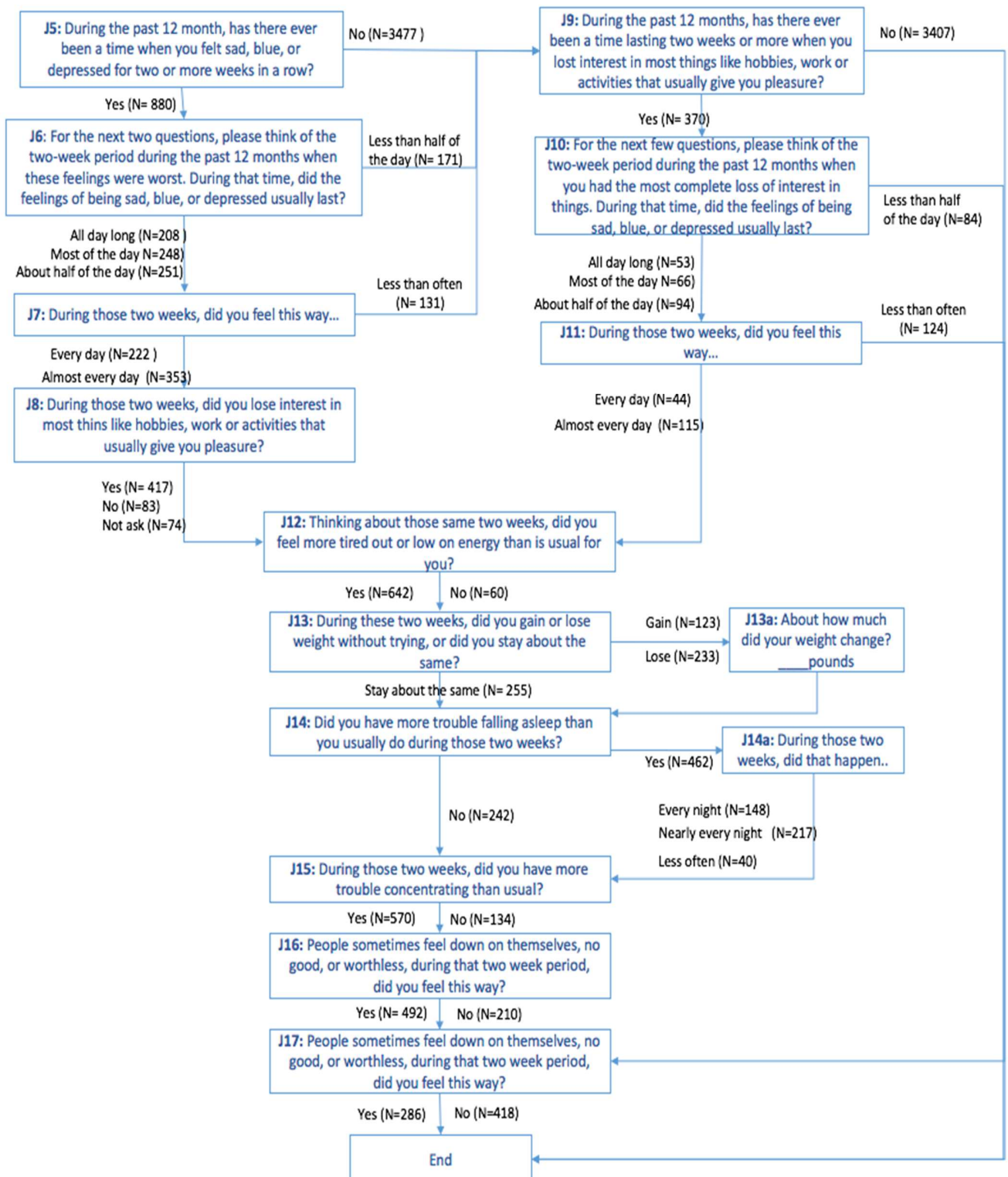
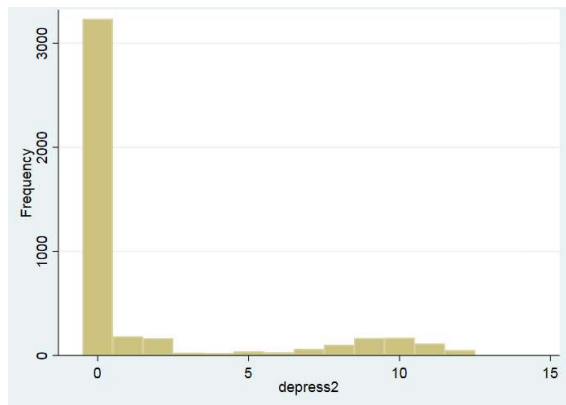
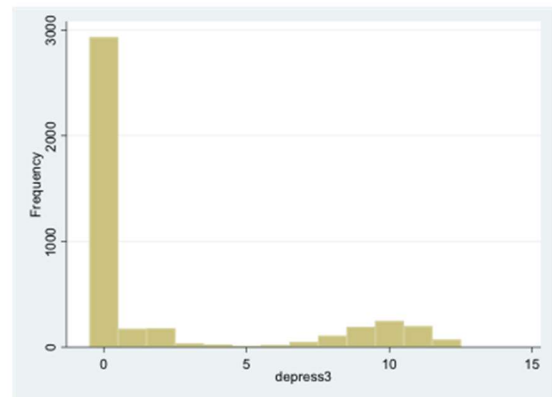


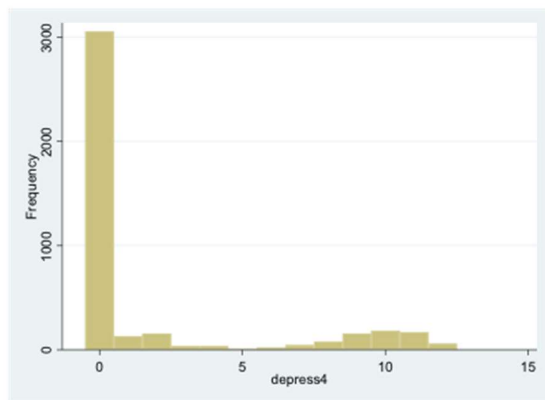
Figure 6. CIDI-SF Questions and Frequency on Wave 2.



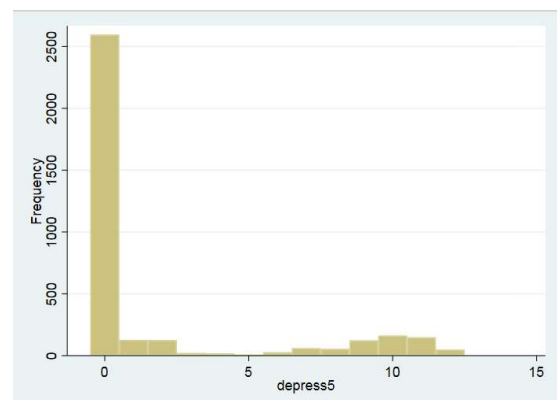
(a) Wave 2



(b) Wave 3



(c) Wave 4



(d) Wave 5

Figure 7. Histogram plots on the frequency of mothers' depression severity.

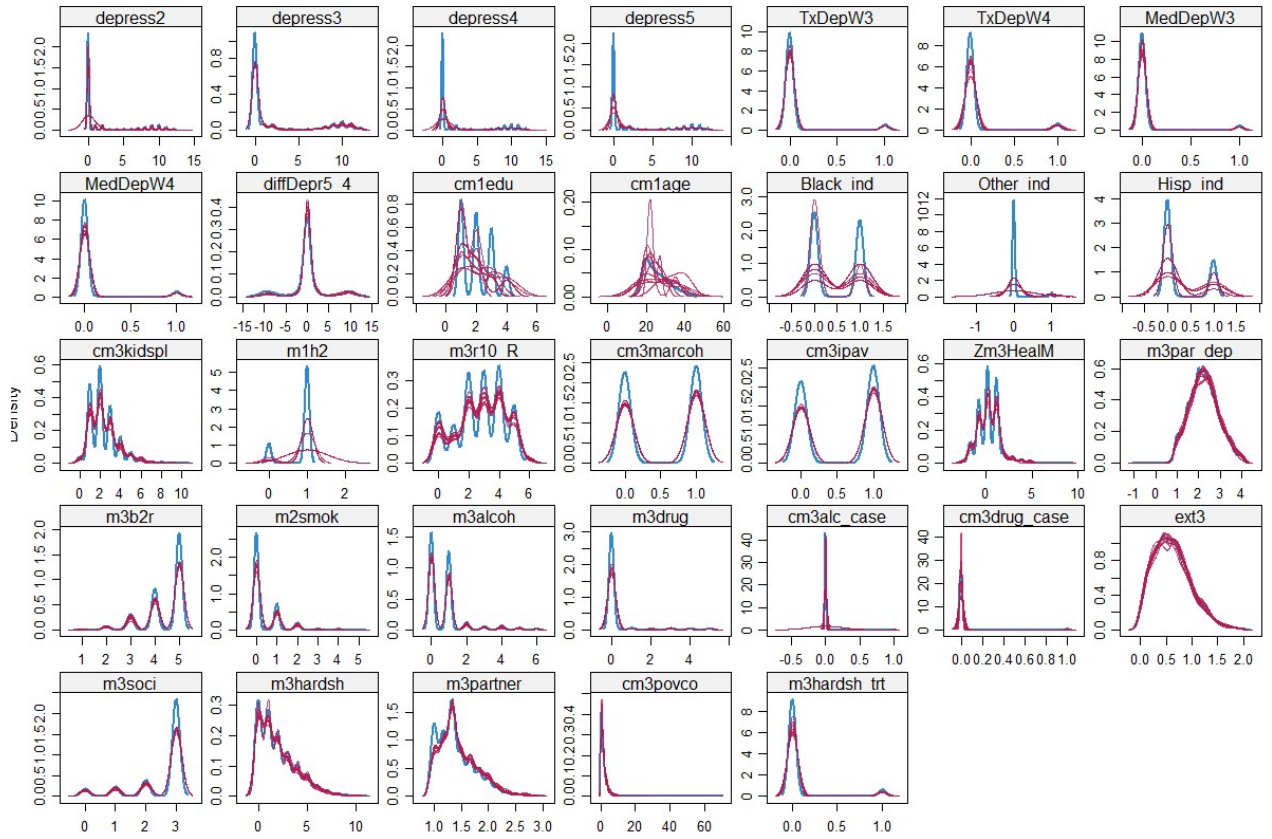


Figure 8. Density Plots for Imputed and Original Data. Depression variables: Depress2 = mothers' depression at Wave 2, depress3 = mothers' depression at Wave 3, depress4 = mothers' depression at Wave 4, depress5 = mothers' depression at Wave 5, diffDepr5_4 = difference from depression on Wave 4 to depression on Wave 5. Treatment: TxDepW3 = psychological treatment at Wave 3, TxdepW4 = psychological treatment at Wave 4, MedDepW3 = medication treatment at Wave 3, MedDepW4 = medication treatment at Wave 4. Covariates at baseline: cm1edu = mothers' education level, cm1age = mothers' age, m1h2 = mothers' foreign born, Black_ind = dummy code on whether mothers' racial is Black, His-ind = dummy code on whether mothers' racial is Hispanic, Other_ind = dummy code on whether mothers' racial is others. Covariates at Wave 2: m2smok = mother smoke. Covariates at Wave 3: cm3kidspl = number of child in mothers' house, m3r1_R = child's health condition, cm3marcoh = mother's cohabitation status, cm3ipav = intimate partner domestic violence, Zm3HealM = mothers' health condition, m3par_dep = parental stress, m3b2r = religious attendance, m3alcoh = alcohol used, cm3alc_case = whether alcohol dependence meet the CIDI-SF criteria, m3drug = drug used, cm3drug_case = whether drug dependence meet the CIDI-SF criteria, ext3 = child's externalizing problems, m3soci = social support, m3hardsh = fanatical hardship, Cm3povco = poverty ratio, m3partner = intimate partner support, m3hardsh-trt = financial hardship for seeing a doctor.

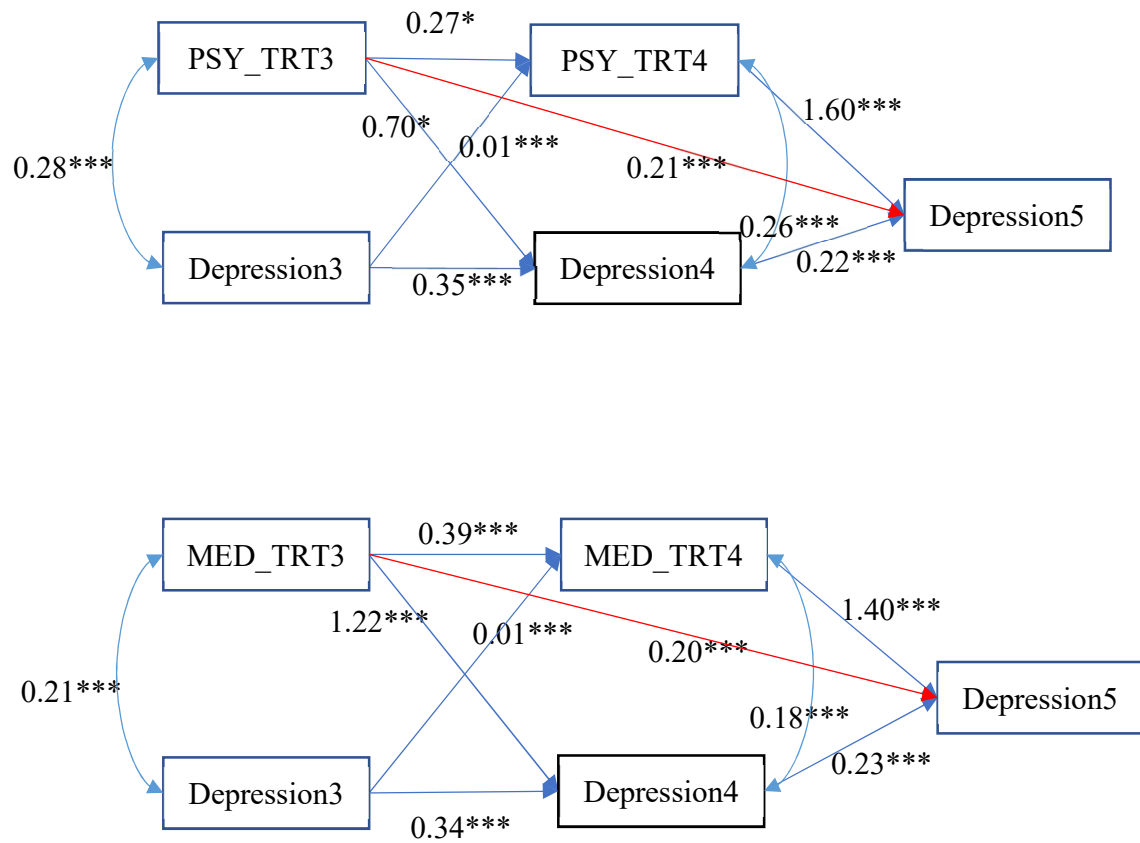


Figure 9. Cross-Lagged Panel Model of Mothers' Depression Featuring Wave 3, 4, and 5. PSY_TRT = psychological treatment. MED-TRT = medication treatment. * $p < .05$. ** $p < .01$. *** $p < .001$.

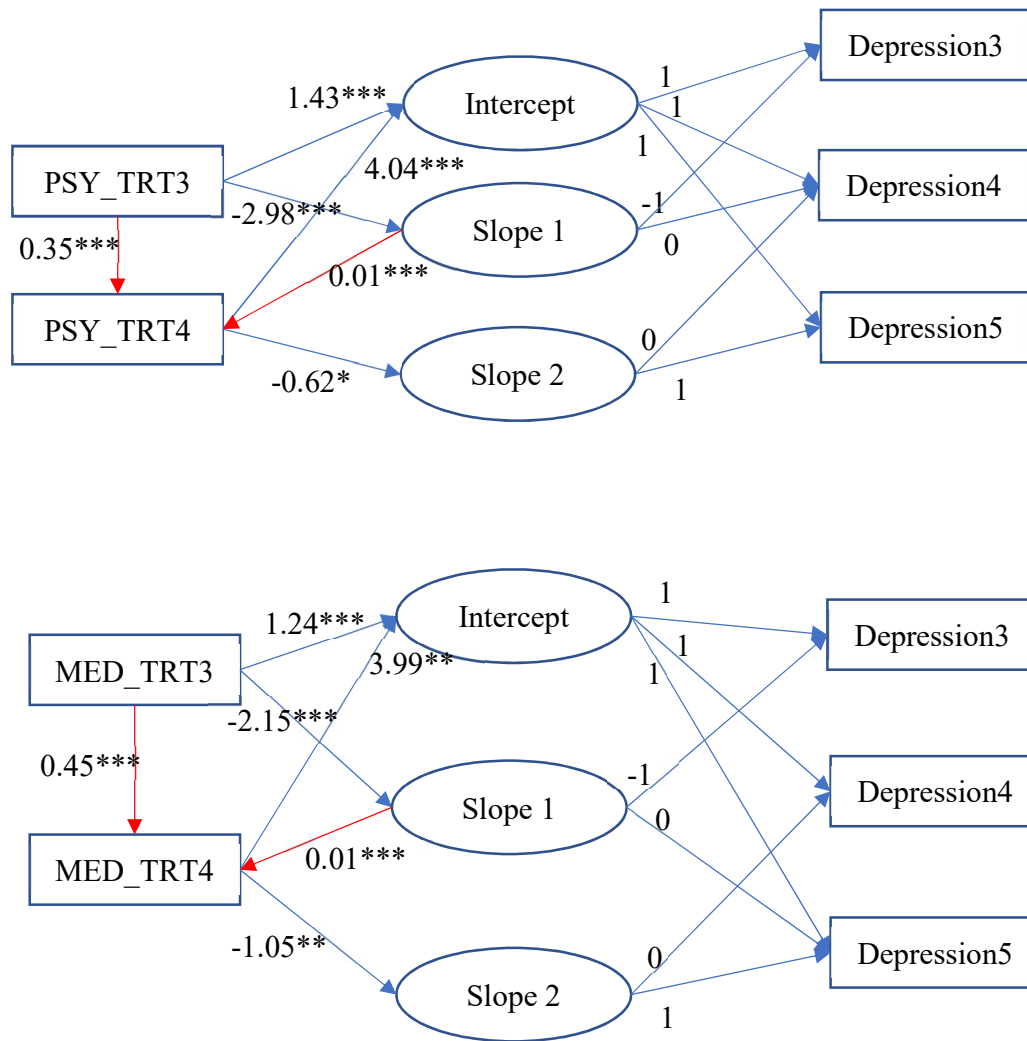


Figure 10. Latent Growth Model of Mothers' Depression Featuring Wave 3, 4, and 5. PSY_TRT = psychological treatment. MED-TRT = medication treatment. . * $p < .05$. ** $p < .01$. *** $p < .001$.

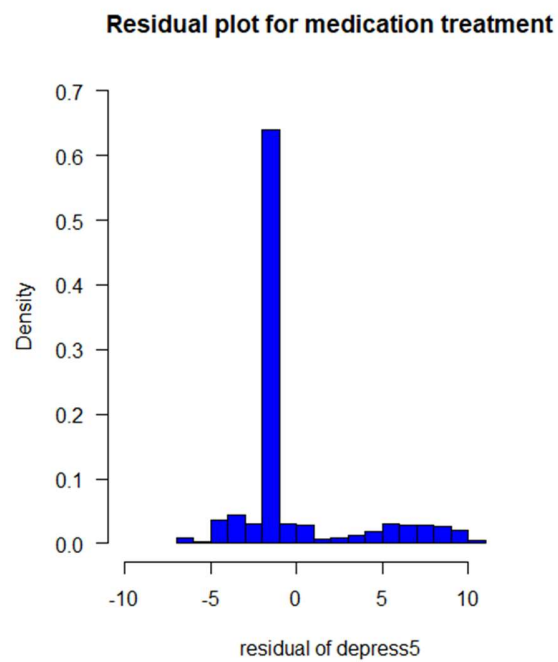
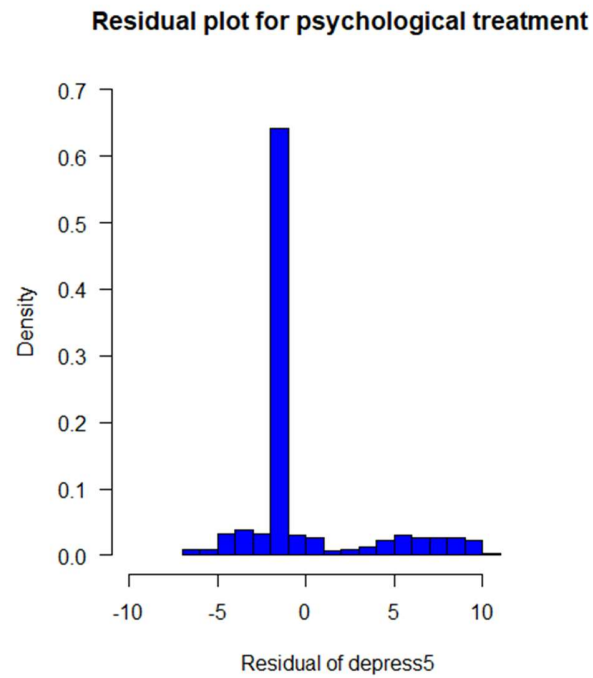


Figure 11. Histogram plots for the distribution of residual of depression at Wave 5 regression on depression on Wave 4 and the two type of treatment when testing research question 3.3.

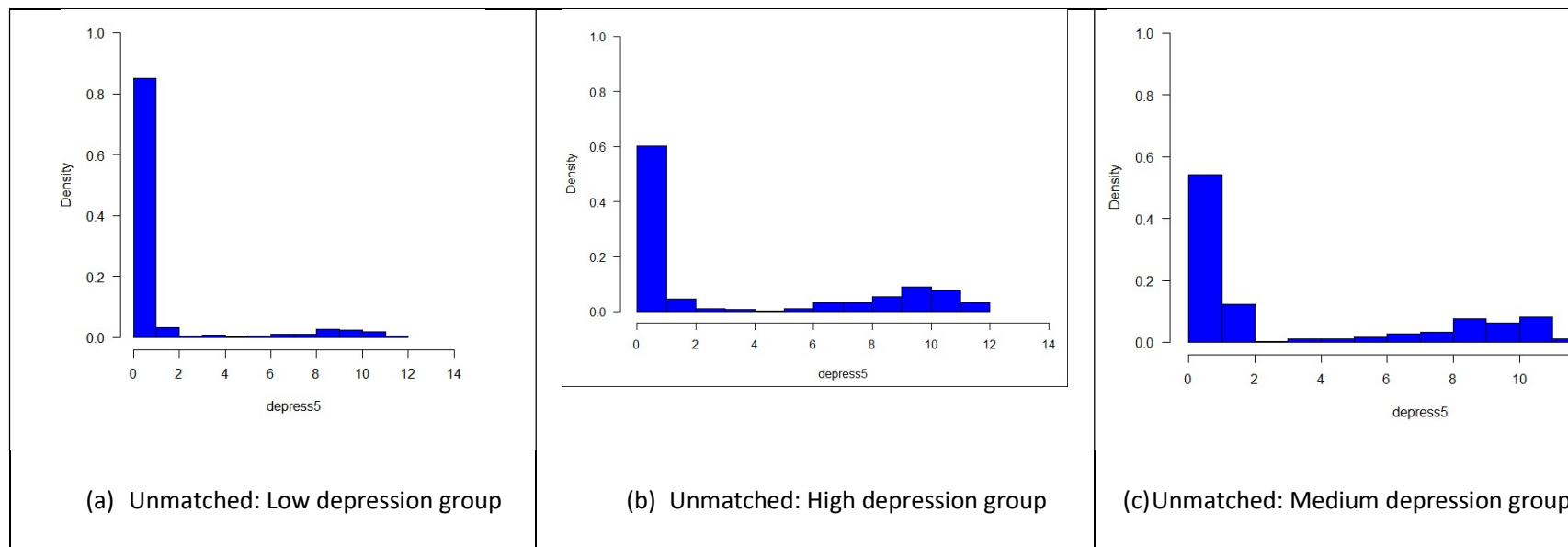


Figure 12. Distribution for mothers' depression at Wave 5 for each unmatched trajectory subgroup.

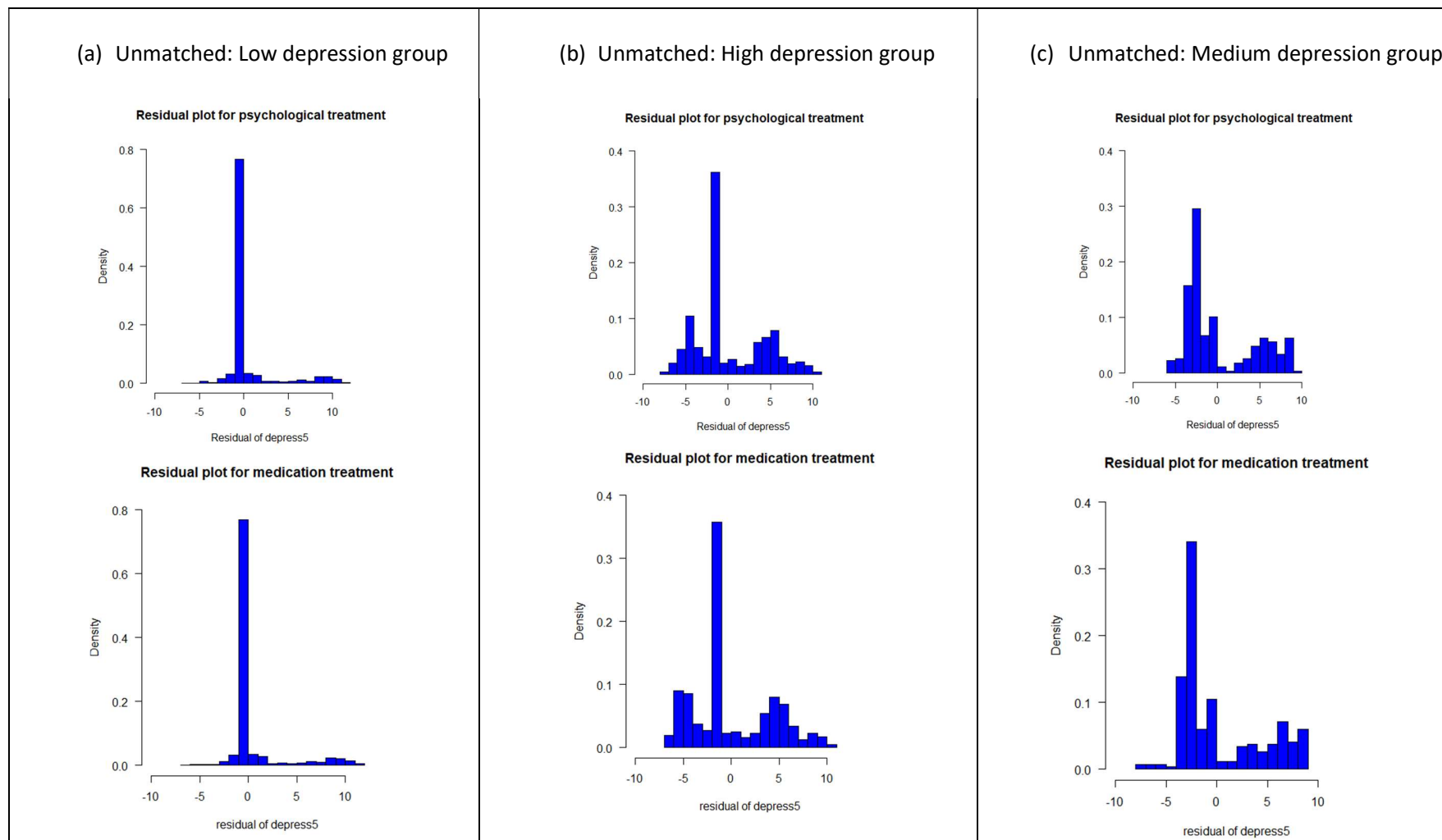


Figure 13. Histogram plot for the distribution of residual of depression at Wave 5 within each trajectory subgroup: regression on Wave 4 depression and the two type of treatment when testing Research Question 3.6.

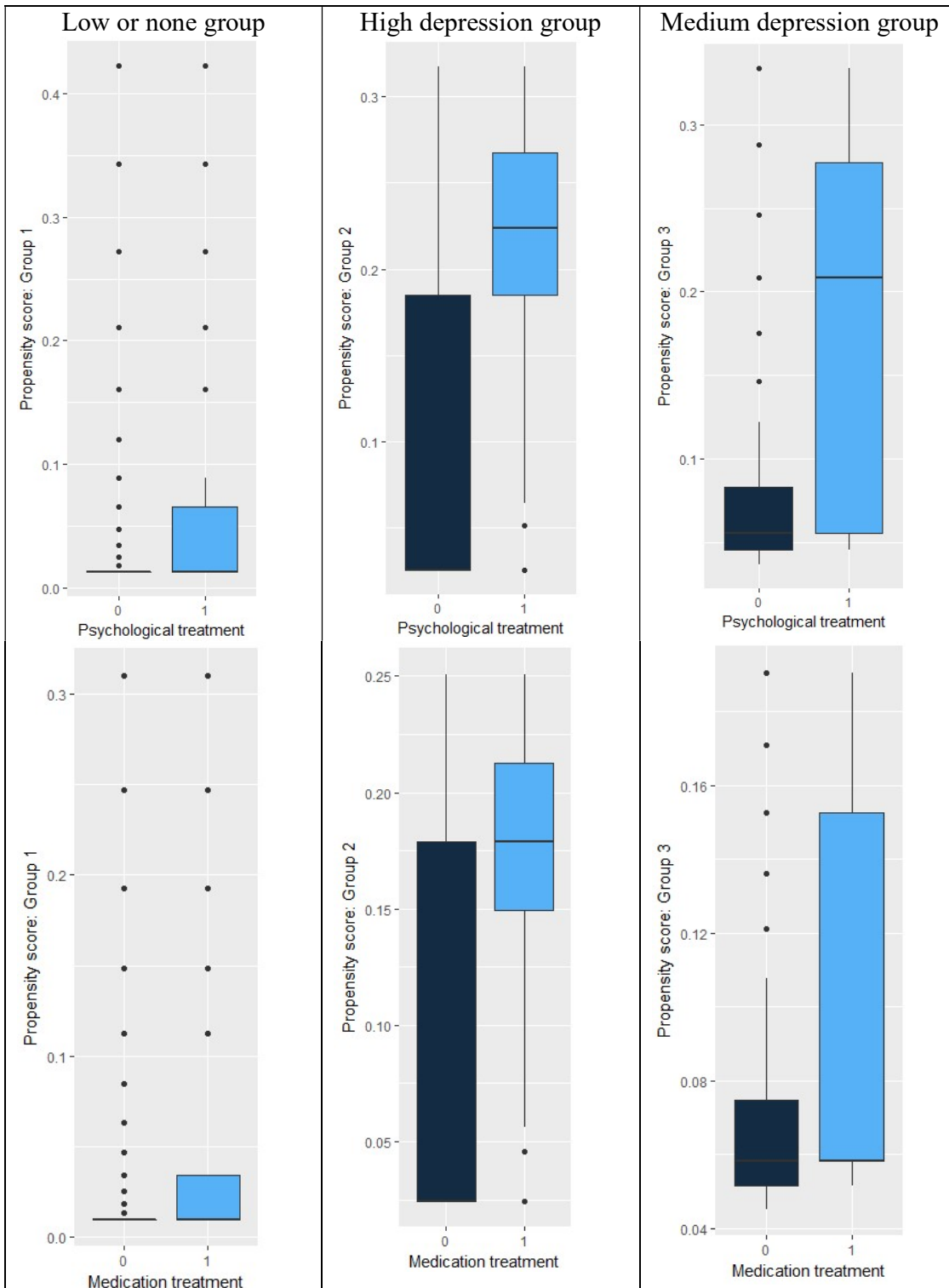


Figure 14. Compare propensity scores between the treatment and control group using unmatched sample: propensity score calculated based on pretest scores only.

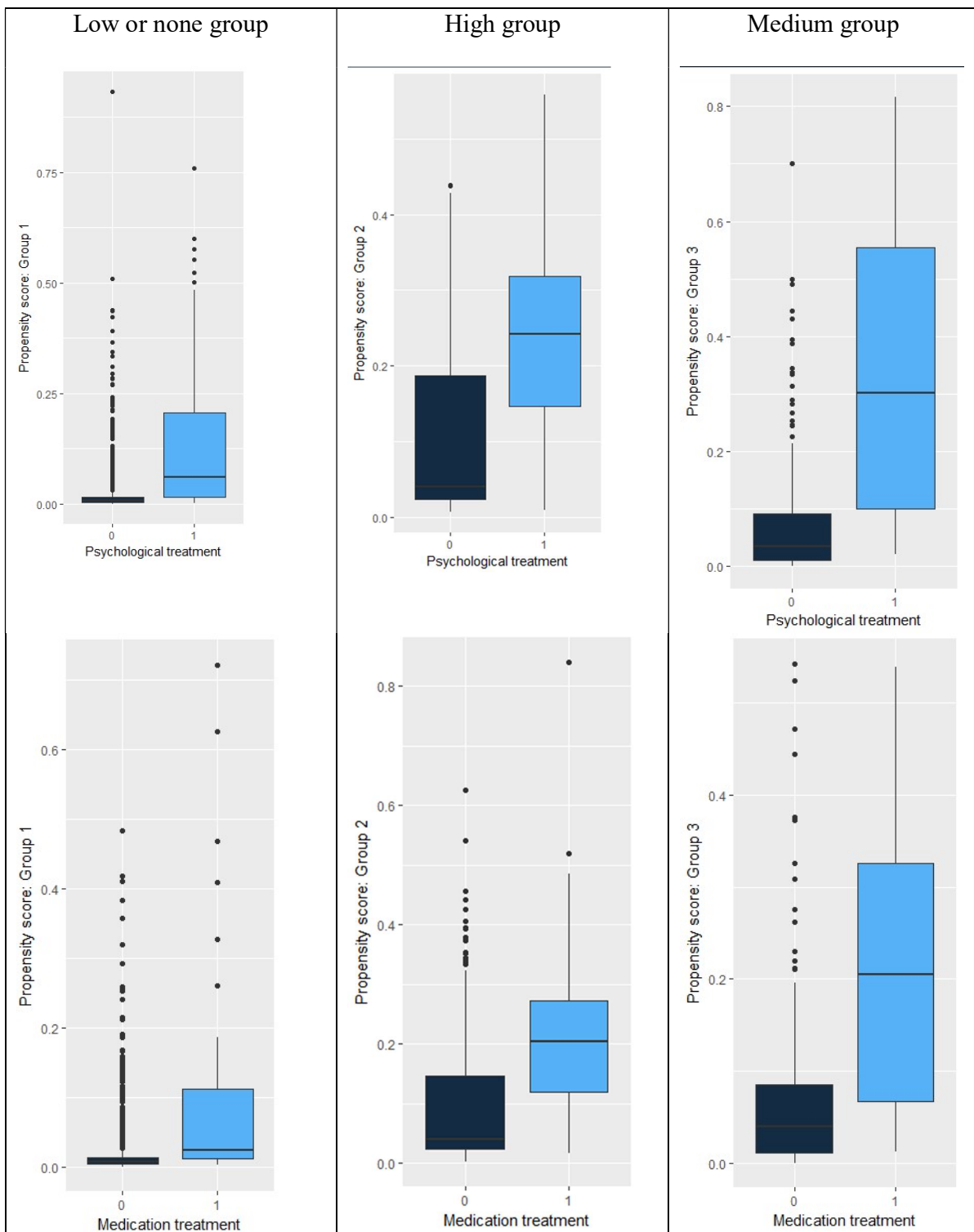


Figure 15. Compare propensity scores between the treatment and control group using unmatched sample: propensity score calculated based on multiple covariates.

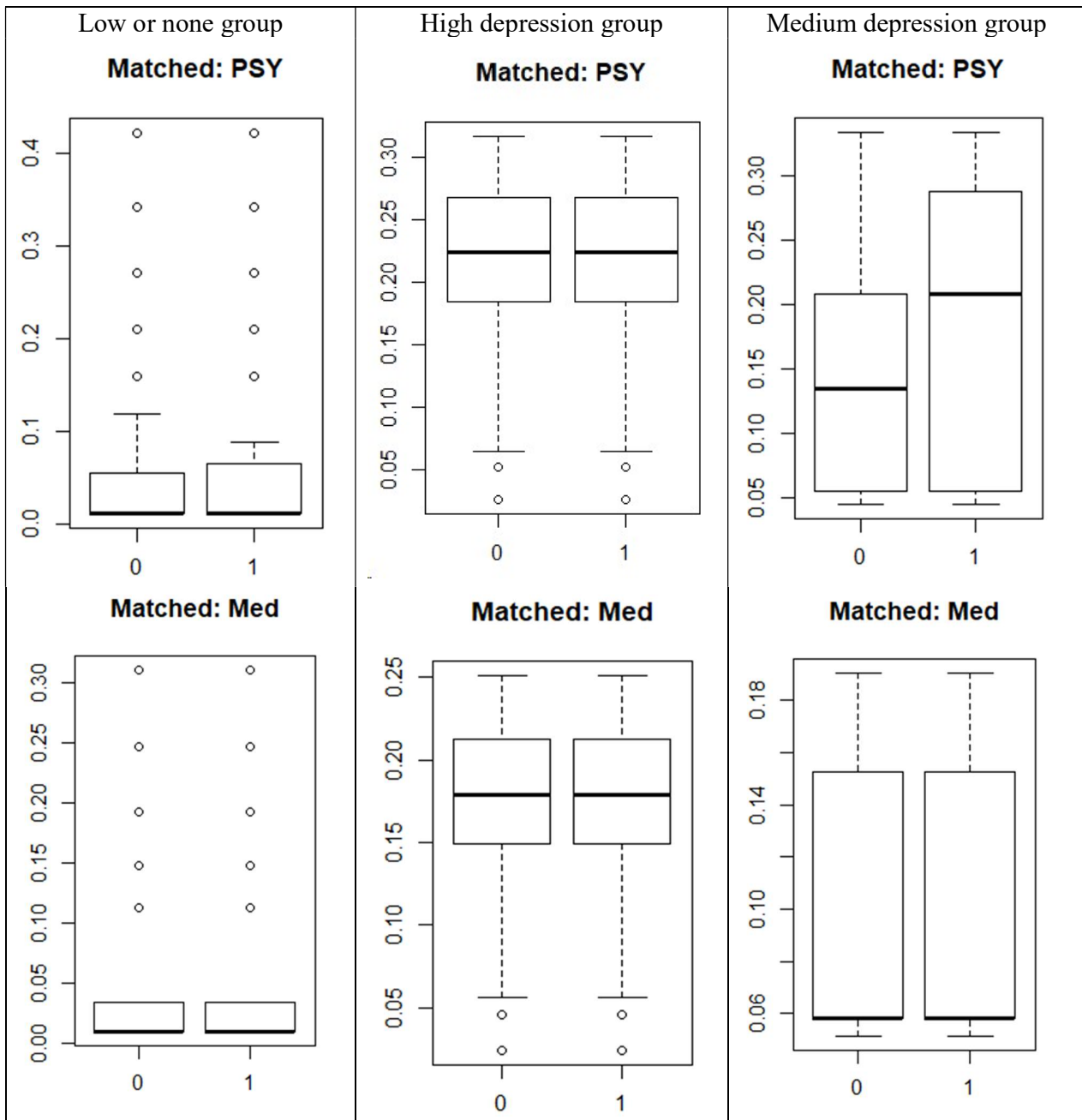


Figure 16. Compare propensity scores between the treatment and control group using matched sample: propensity score calculated based on pretest score only.

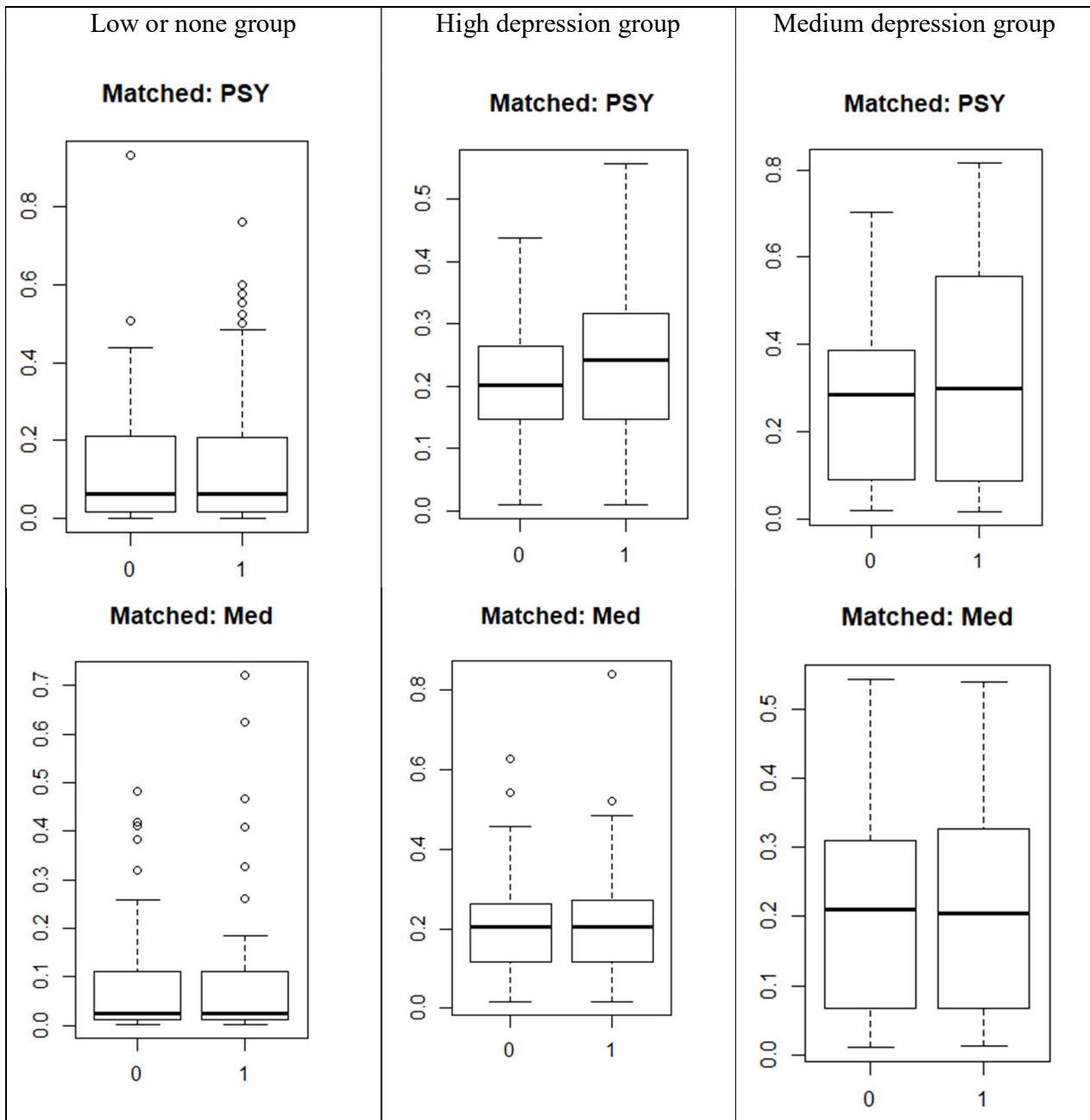


Figure 17. Compare propensity scores between the treatment and control group using matched sample: propensity score calculated based on multiple covariates.

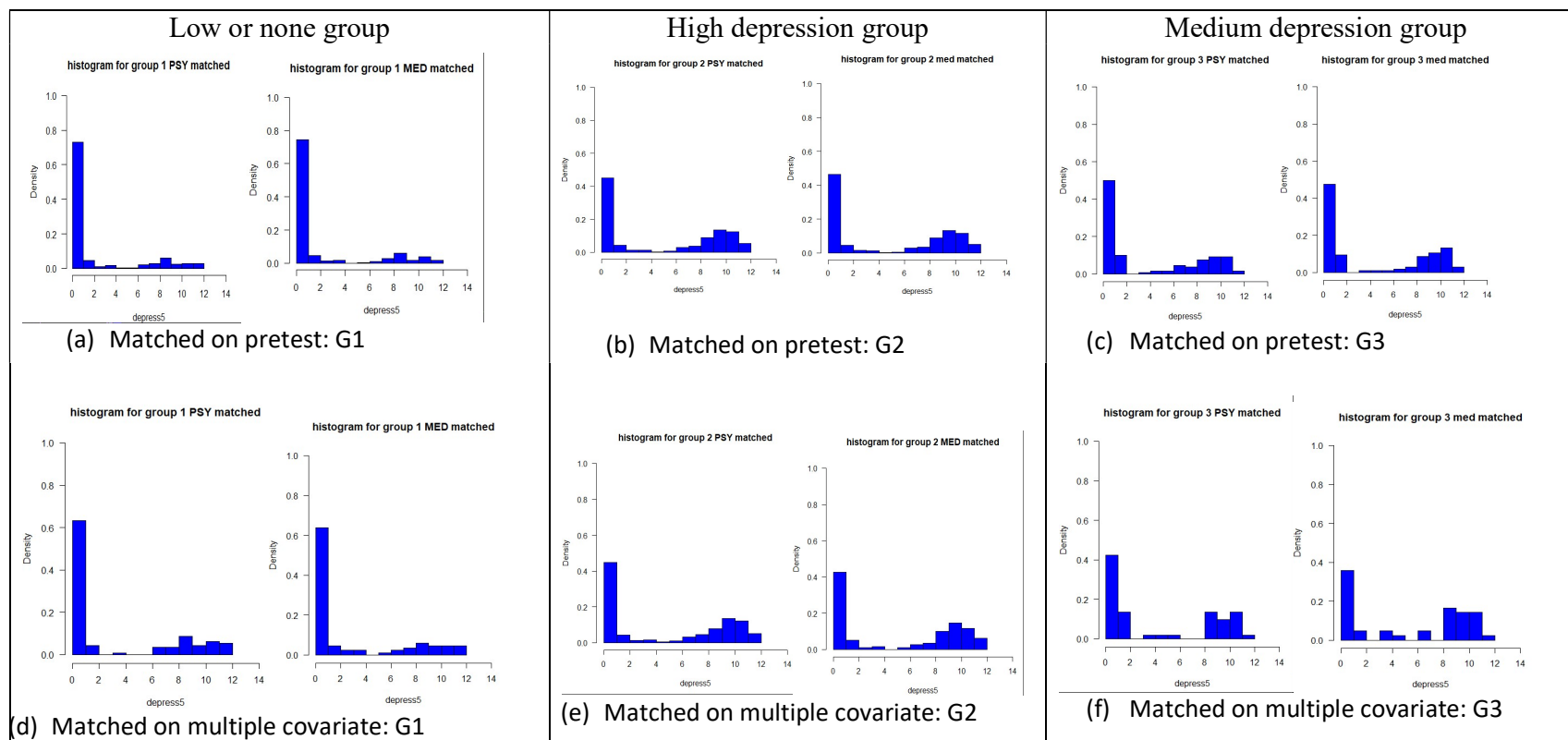


Figure 18. Distribution for mothers' depression at Wave 5 for matched samples within each trajectory group.

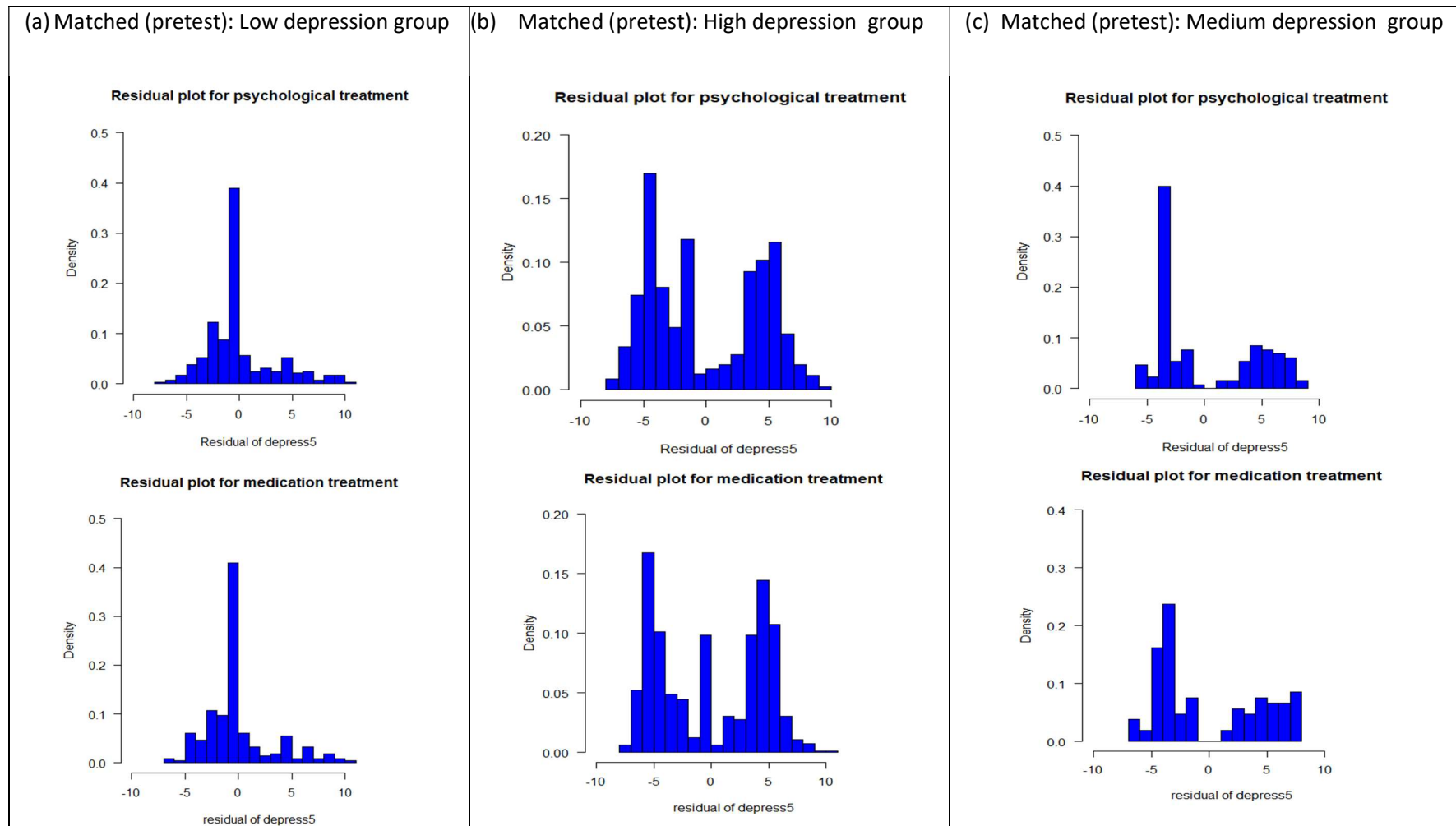


Figure 19. Histogram Plots for the distribution of residual of depression at Wave 5 within each trajectory subgroup using matched samples based on pretest outcome as the only covariate: regression on Wave 4 depression and the two type of treatment.

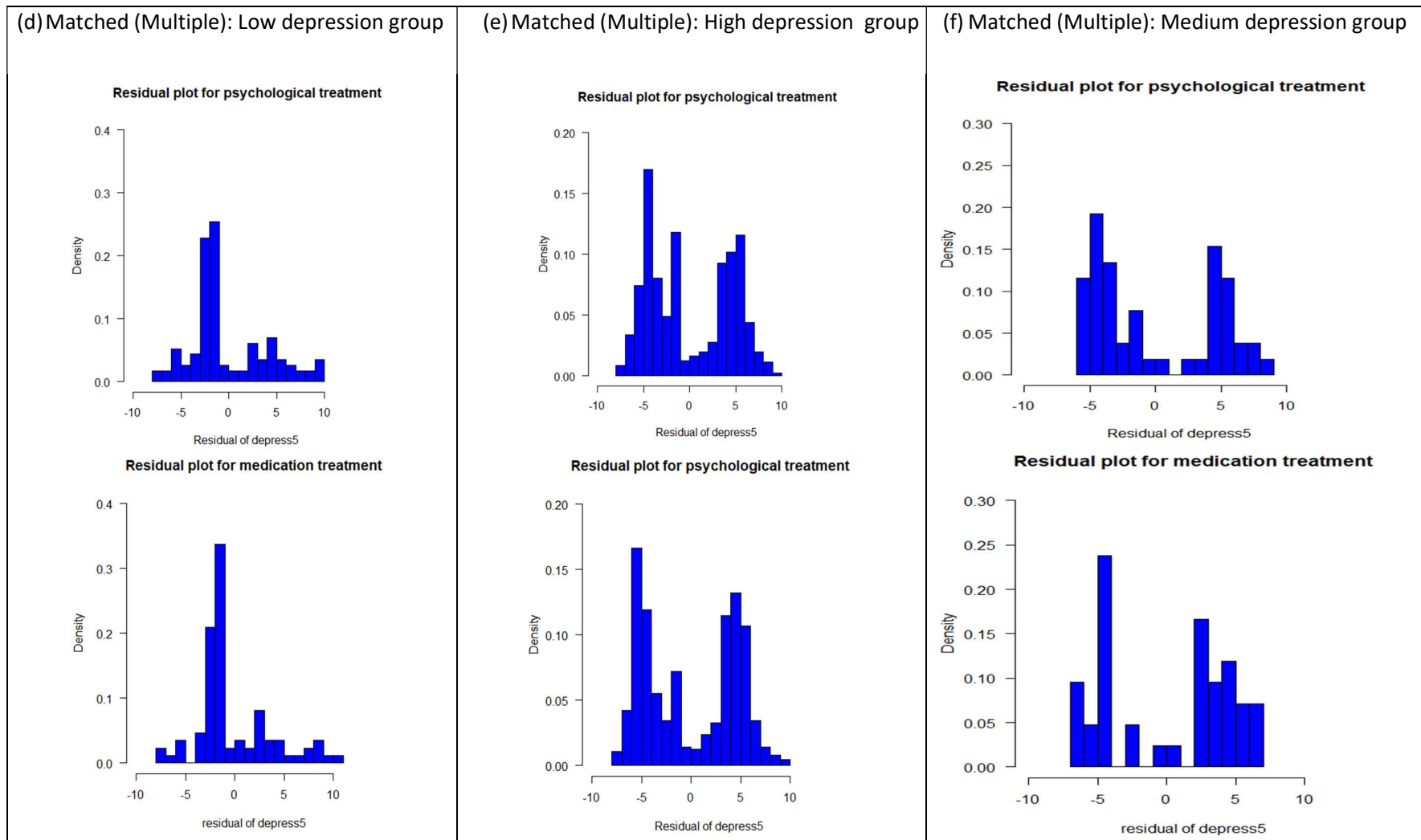


Figure 20. Histogram plots for the distribution of residual of depression at wave 5 within each trajectory subgroup using matched samples based on multiple covariate: regression on Wave 4 depression and the two type of treatment.

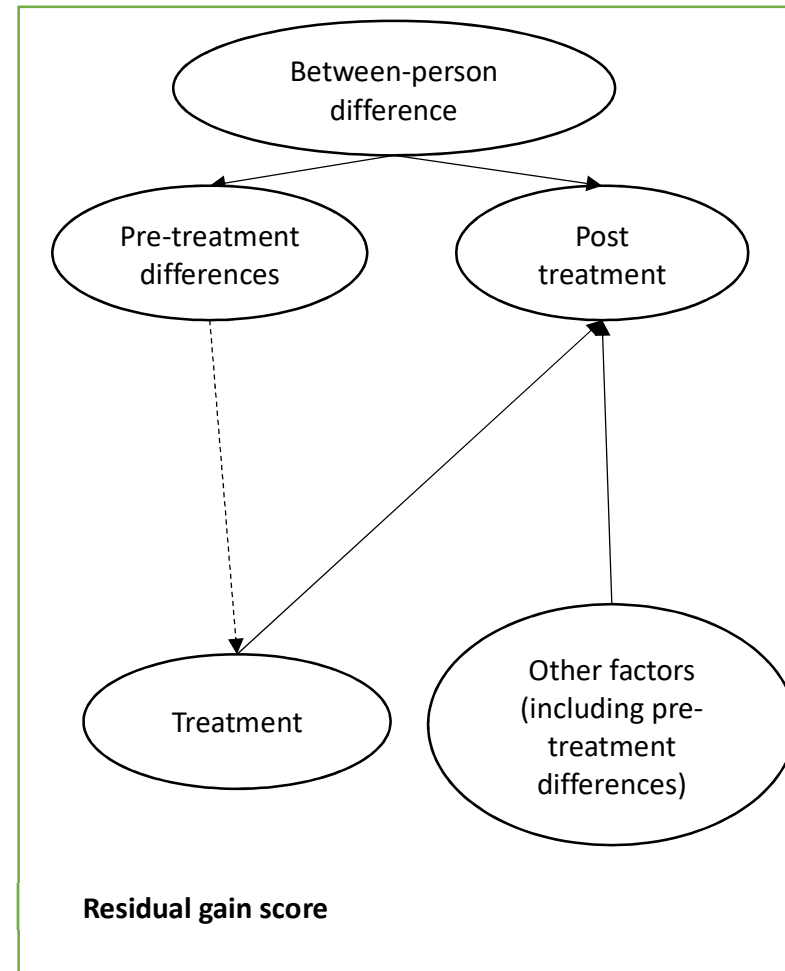
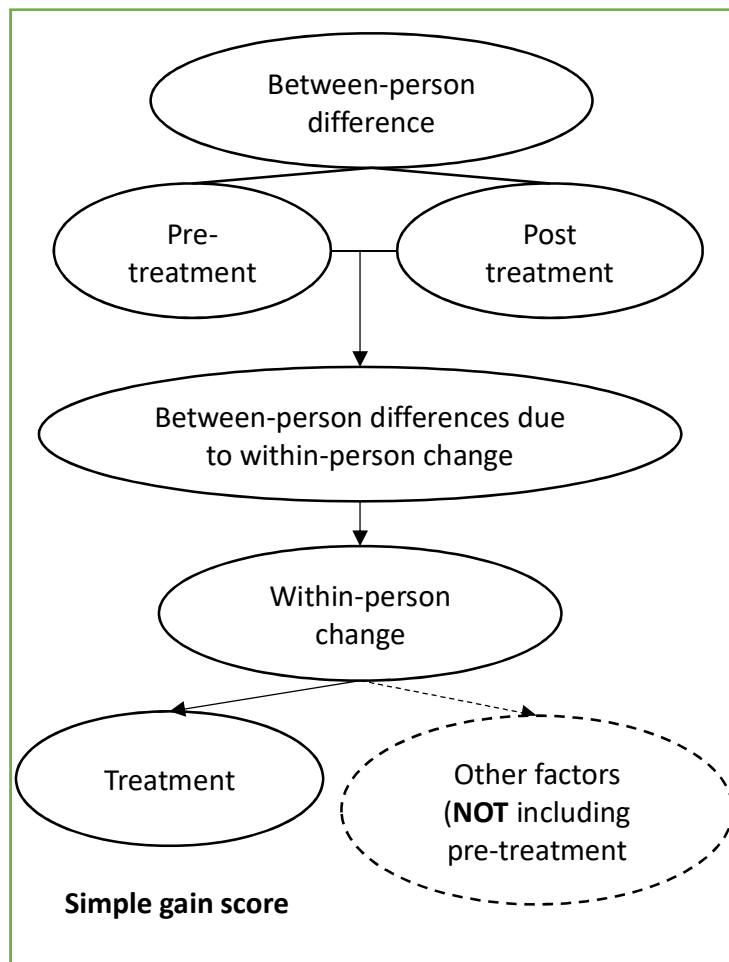


Figure 21. Simple Gain Score and Residual Gain Score.

APPENDIX A

R Syntax for Simulating Lord's Paradox and Violated Assumptions

*# Please install the package first by running the first line, then run the rest of the syntax.
After you run all the syntax, the result will be in the folder: D:/paradox*

```
install.packages("mvtnorm")  # install package

rm(list=ls()) # remove everything
output <- matrix(nrow = 100, ncol = 26)
colnames(output) <-
c("id_data", "back", "NN", "NT", "NG", "NB", "M0g", "M0b", "M1g", "M1b",
  "SDg", "SDb", "cor_g10", "cor_b10", "y0g_mean",
  "y0b_mean", "y1g_mean", "y1b_mean", "y10g_mean",
  "y10b_mean", "diff_diff", "t_test", "t_pv", "reg_girl_coef",
  "reg_girl_ttest", "reg_girl_pv")

d <- 1
NT <- 1000 # sample size
SDb <- 15  # standard deviation for male
M1g <- 130 # Setting females' posttest mean
M1b <- 160 # Setting males' posttest mean
M0 <- 145  # Setting pretest grand mean
M1 <- 145  # Setting posttest grand mean
NG <- 500  # sample size for female
NB <- NT - NG # sample size for males
for (M0g in c(130, 145)) { # pretest mean for females: two variations
  for (M0b in c(145, 160)) { # Pretest mean for males: two variations
    for (back in 0:1) {
      for (NN in 0:1) { # test distribution: NN=1(=Not normal), NN=0(normal)
        for (SDg in c(5, 15)) { # pretest SD for females: two variations
          for (cor_b10 in c(0, 0.48, 0.96)) { # correlation between pretest and posttest for
            females
```

```

    for (cor_g10 in c(0,0.48, 0.96)){ # correlation between pretest and posttest
for males
        if (M0g==130 & M0b==160) {nstop=1
        } else if (M0g==145 & M0b==145) {nstop=1 #pre= post=! ∨ 30
        } else {nstop = 0}
        if (nstop == 1) {
            if (cor_b10==0.48 & cor_g10 == 0.48) {nstop=1 #pre=! post=!
            } else if (back==0 & cor_b10==0.96 & cor_g10 == 0) {nstop=1
            } else if (back==0 & cor_b10==0 & cor_g10 == 0.96) {nstop=1
            } else if (back==0 & cor_b10==0.48 & cor_g10 == 0) {nstop=1
            } else if (back==0 & cor_b10==0 & cor_g10 == 0.48) {nstop=1
            } else if (back==1 & cor_b10==0.48 & cor_g10 == 0.48) {nstop=1
            } else {stopr=0}
            if (stopr==1){ # don't run the situation that SDb=SDg=5,or
correlation=0
                set.seed(123456)
                N <- 1000 # simulate 1000 times
                y0g_mean <- rep(NA,N) # store the simulated females' pretest means
                y1g_mean <- rep(NA,N) # store the simulated females' posttest means
                y0b_mean <- rep(NA,N) # store the simulated males' posttest means
                y1b_mean <- rep(NA,N) # store the simulated males' posttest means
                y10g_mean <- rep(NA, N) # store the females' post and pretest mean
difference
                y10b_mean <- rep(NA, N) # store the males' post and pretest mean
difference
                diff_diff <- rep(NA, N) #mean difference of difference
                t_test <- rep(NA, N) # store the t_test parameter
                t_pv <- rep(NA, N) # store the t_test's p_value
                reg_girl_coef <- rep(NA, N) # regression group effect
                reg_girl_ttest <- rep(NA, N) # t_test for group effect
                reg_girl_pv <- rep(NA, N) # p-valut for group effect

                #simulation for N=1000 times
                for(i in 1:N) {

                    girl <- rep(NA, NT)
                    girl[1:NG] <- 1
                    girl [NG+1:NB] <- 0
                    y0 <- rep(NA, NT)
                    y1 <- rep(NA, NT)

                    # run Lord's backward paradox
                    if (back==1){
                    y0g <- rnorm(NG, M0g, SDg) # generate data for females

```

```

y0b <- rnorm(NB, M0b, SDb) #generate data for males
y1g <- rep(NA, NG)
y1b <- rep(NA, NB)
SD <- 15
cor <- cor_g10 #the whole sample us the same coefficient for pretest
e<- sqrt(SD^2 - cor_g10^2*SD^2) # error in the regression fomula for
females
a0<- M1-cor_g10*M0 # intercept in the regression fomula for males
e_i <- rnorm (N, 0, e)
for (j in 1:NG){
  y0[j] <- y0g[j]
}
for (k in 1:NB){
  y0[k+NG] <- y0b[k]
}
y1=a0+cor*y0+0*girl+e_i

for (j in 1:NG){
  y1g[j] <- y1[j]
}
for (k in 1:NB){
  y1b[k] <- y1[k+NG]
}}

# run Lord's original paradox
else if (back==0) {
library(mvtnorm)
varg <- matrix(c(SDg^2,cov_g10 <-
cor_g10*sqrt(SDg^2*SDg^2),cov_g10,SDg^2),ncol = 2)
varb <- matrix(c(SDb^2,cov_b10 <-
cor_b10*sqrt(SDb^2*SDb^2),cov_b10,SDb^2),ncol = 2)
g <- rmvnorm(n=NG, mean=c(M0g,M1g),sigma =varg) # generate data
for females
b <- rmvnorm(n=NB, mean=c(M0b,M1b),sigma =varb) #generate data
for males

y0g <- g[,1]
y1g <- g[,2]
y0b <- b[,1]
y1b <- b[,2]
}

# Generate non_normal distribution data
if (NN==1){
m <- 1

```

```

while (m < NG+1) {
  if (round(m/4) != (m/4)) {
    y0g[m] <- 70
  }
  if (round(m/5) != (m/5)) {
    y1g[m] <- 70
  }
  m=m+1
}

p <- 1
while (p < NB+1) {
  if (round(p/4) != (p/4)) {
    y0b[p] <- 70
  }
  if (round(p/5) != (p/5)) {
    y1b[p] <- 70
  }
  p=p+1
}
}

y10g <- y1g - y0g      # calculate the difference of females
y10b <- y1b - y0b      # calculate the difference of males
y0g_mean[i] <- mean(y0g)
y0b_mean[i] <- mean(y0b)
y1g_mean[i] <- mean(y1g)
y1b_mean[i] <- mean(y1b)
y10g_mean[i] <- mean(y10g)
y10b_mean[i] <- mean(y10b)
diff_diff[i] <- mean(y10g) - mean(y10b)
diff_gap[i] <- abs(mean(y0b-y0g))-abs(mean(y1b-y1g))

for (j in 1:NG){
  y0[j] <- y0g[j]
  y1[j] <- y1g[j]
}
for (k in 1:NB){
  y0[k+NG] <- y0b[k]
  y1[k+NG] <- y1b[k]
}

# for regression analysis
reg <- lm(y1 ~ y0 + girl)
reg_girl_coef[i] <- reg$coefficients[3]
reg_girl_ttest[i] <- coef(summary(reg)) [3,3]
reg_girl_pv[i] <- coef(summary(reg)) [3,4]

```

```

# For t-test analysis
y10 <- y1-y0
tt<- t.test(y10 ~ girl) # test test
t_test[i] <- tt$statistic
t_pv[i] <- tt$p.value
}

# give output and store
result_data <- data.frame(y0g_mean,y0b_mean, y1g_mean, y1b_mean,
y10g_mean,
                        y10b_mean, diff_diff, t_test, t_pv, reg_girl_coef,
                        reg_girl_ttest, reg_girl_pv)
result_mean <- c(colMeans(result_data, dims = 1))
parameter <- matrix
(c(d,back,NN,NT,NG,NB,M0g,M0b,M1g,M1b,SDg,SDb,cor_g10,cor_b10),
  nrow = 1,ncol = 14)
outp<- c (parameter,result_mean)
output[d,]<- outp
assign(paste("data",d),result_data)

# set the data file location in D: and create a data file called "paradox"
if (dir.exists("C:/Users/LHEVA/Dropbox/3.Share folder/FF
data/Analysis/Dissertation/Simulation/R/MMCC")) {
  setwd(dir = "C:/Users/LHEVA/Dropbox/3.Share folder/FF
data/Analysis/Dissertation/Simulation/R/MMCC")
} else {
  dir.create("C:/Users/LHEVA/Dropbox/3.Share folder/FF
data/Analysis/Dissertation/Simulation/R/MMCC")
  setwd(dir = "C:/Users/LHEVA/Dropbox/3.Share folder/FF
data/Analysis/Dissertation/Simulation/R/MMCC")
}

# set the data file location in D: and create a data file called "paradox"
if (dir.exists("D:/paradox")) {
  setwd(dir = " D:/paradox ")
} else {
  dir.create("D:/paradox ")
  setwd(dir = " D:/paradox ")
}

# save the simulated data file and result file in current location
filename <- paste("data",d,".txt",sep = "")
write.table(result_data,
filename,row.names=FALSE,sep="\t",quote=FALSE)
write.csv(output,"results.csv",row.names=FALSE)
d=d+1}}}}}}}}

```


APPENDIX B

R Syntax for Testing Matched and Unmatched Simulated Load's Paradox and Reversed Lord's Paradox

```
rm(list=ls()) # remove everything
# total sample size is 1000
output <- matrix(nrow = 4, ncol = 64)
colnames(output) <-
c("id_data", "back", "NN", "NT", "NG", "NB", "M0g", "M0b", "M1g", "M1b",
  "SDg", "Sdb", "corr", "y0g_mean",
  "y0b_mean", "y1g_mean", "y1b_mean", "y10g_mean",
  "y10b_mean", "or_diff", "or_t", "or_tpv", "or_reg_b",
  "or_reg_t", "or_reg_tpv", "or_up_diff", "or_up_t", "or_up_tpv",
  "or_up_reg_b",
  "or_up_reg_t", "or_up_reg_tpv", "or_low_diff", "or_low_t", "or_low_tpv",
  "or_low_reg_b",
  "or_low_reg_t", "or_low_reg_tpv", "or_up_p.diff", "or_up_p.t",
  "or_up_p.tpv", "or_up_reg_p.b",
  "or_up_reg_p.t", "or_up_p.reg_tpv", "or_low_p.diff", "or_low_p.t",
  "or_low_p.tpv", "or_low_p.reg_b",
  "or_low_p.reg_t", "or_low_p.reg_tpv", "or_y0_diff", "or_y0_t",
  "or_y0_tpv",
  "or_upy0_p.diff", "or_upy0_p.t",
  "or_upy0_p.tpv", "or_upy0_diff", "or_upy0_t", "or_upy0_tpv",
  "or_lowy0_p.diff", "or_lowy0_p.t", "or_lowy0_p.tpv", "or_lowy0_diff",
  "or_lowy0_t", "or_lowy0_tpv")

d <- 1
NT <- 1000 # sample size
SDg <- 15
Sdb <- 15
```

```

M0g <- 130
M0b <- 160
M1g <- 130
M1b <- 160
M0 <- 145      # overall pretest mean
M1 <- 145      # overall pretest mean
NG <- 500
NB <- NT - NG
corr <- -0.48
NN <- 0

for (back in 0:1){
  set.seed(123456)
  N <- 1000 # simulate 1000 times
  y0g_mean <- rep(NA,N) #females' pretest means
  y1g_mean <- rep(NA,N) #females' posttest means
  y0b_mean <- rep(NA,N) #males' posttest means
  y1b_mean <- rep(NA,N) #males' posttest means
  y10g_mean <- rep(NA, N) #females' post pretest mean difference
  y10b_mean <- rep(NA, N) #females' post pretest mean difference

  or_diff <- rep(NA, N) #mean difference if difference
  or_t <- rep(NA, N) #t_test parameter
  or_tpv <- rep(NA, N) #t_test p_value

  or_reg_b <- rep(NA, N) # regression group effect
  or_reg_t <- rep(NA, N) # t_test for group effect
  or_reg_tpv <- rep(NA, N) # p-valut for group effect

  or_up_diff <- rep(NA, N) #mean difference if difference
  or_up_t <- rep(NA, N) #t_test parameter
  or_up_tpv <- rep(NA, N) #t_test p_value

  or_up_reg_b <- rep(NA, N) # regression group effect
  or_up_reg_t <- rep(NA, N) # t_test for group effect
  or_up_reg_tpv <- rep(NA, N) # p-valut for group effect

  or_low_diff <- rep(NA, N) #mean difference if difference
  or_low_t <- rep(NA, N) #t_test parameter
  or_low_tpv <- rep(NA, N) #t_test p_value

  or_low_reg_b <- rep(NA, N) # regression group effect
  or_low_reg_t <- rep(NA, N) # t_test for group effect
  or_low_reg_tpv <- rep(NA, N) # p-valut for group effect

  or_up_p.diff <- rep(NA, N) #mean difference if difference

```

```

or_up_p.t <- rep(NA, N) #t_test parameter
or_up_p.tpv <- rep(NA, N) #t_test p_value

or_up_p.reg_b <- rep(NA, N) # regression group effect
or_up_p.reg_t <- rep(NA, N) # t_test for group effect
or_up_p.reg_tpv <- rep(NA, N) # p-valut for group effect

or_low_p.diff <- rep(NA, N) #mean difference if difference
or_low_p.t <- rep(NA, N) #t_test parameter
or_low_p.tpv <- rep(NA, N) #t_test p_value

or_low_p.reg_b <- rep(NA, N) # regression group effect
or_low_p.reg_t <- rep(NA, N) # t_test for group effect
or_low_p.reg_tpv <- rep(NA, N) # p-valut for group effect

or_upy0_p.diff <- rep(NA, N)
or_upy0_p.t <- rep(NA, N)
or_upy0_p.tpv <- rep(NA, N)

or_upy0_diff <- rep(NA, N)
or_upy0_t <- rep(NA, N)
or_upy0_tpv <- rep(NA, N)

or_lowy0_p.diff <- rep(NA, N)
or_lowy0_p.t <- rep(NA, N)
or_lowy0_p.tpv <- rep(NA, N)

or_lowy0_diff <- rep(NA, N)
or_lowy0_t <- rep(NA, N)
or_lowy0_tpv <- rep(NA, N)

or_lowy0_p.diff <- rep(NA, N)
or_lowy0_p.t <- rep(NA, N)
or_lowy0_p.tpv <- rep(NA, N)

or_y0_diff <- rep(NA, N)
or_y0_t <- rep(NA, N)
or_y0_tpv <- rep(NA, N)

#simulation for N=1000 times
for(i in 1:N) {

  girl <- rep(NA, NT)
  girl[1:NG] <- 1
  girl [NG+1:NB] <- 0
  y0 <- rep(NA, NT)

```

```

y1 <- rep(NA, NT)

# run Lord's backward paradox
if (back==1){
  y0g <- rnorm(NG, M0g, SDg) # generate data for females
  y0b <- rnorm(NB, M0b, SDb) #generate data for males
  y1g <- rep(NA, NG)
  y1b <- rep(NA, NB)
  SD <- 15
  e<- sqrt(SD^2 - corr^2*SD^2) # error in the regression fomula for females
  a0<- M1-corr*M0 # intercept in the regression fomula for males
  e_i <- rnorm (N, 0, e)
  for (j in 1:NG){
    y0[j] <- y0g[j]
  }
  for (k in 1:NB){
    y0[k+NG] <- y0b[k]
  }
  y1=a0+corr*y0+0*girl+e_i

  for (j in 1:NG){
    y1g[j] <- y1[j]
  }
  for (k in 1:NB){
    y1b[k] <- y1[k+NG]
  }
}

# run Lord's original paradox
else if (back==0) {
  library(mvtnorm)
  varg <- matrix(c(SDg^2,cov_g10 <-corr*sqrt(SDg^2*SDg^2),cov_g10,SDg^2),ncol
= 2)
  varb <- matrix(c(SDb^2,cov_b10 <-corr*sqrt(SDb^2*SDb^2),cov_b10,SDb^2),ncol
= 2)
  g <- rmvnorm(n=NG, mean=c(M0g,M1g),sigma =varg) # generate data for females
  b <- rmvnorm(n=NB, mean=c(M0b,M1b),sigma =varb) #generate data for males
  y0g <- g[,1]
  y1g <- g[,2]
  y0b <- b[,1]
  y1b <- b[,2]
}

```

```

# Generate non_normal distribution data
if (NN==1){
  m <- 1
  while (m < NG+1) {
    if (round(m/4) != (m/4)){
      y0g[m] <- 70
    }
    if (round(m/4) != (m/4)){
      y1g[m] <- 70
    }
    m=m+1
  }

  p <- 1
  while (p < NB+1) {
    if (round(p/4) != (p/4)){
      y0b[p] <- 70
    }
    if (round(p/4) != (p/4)){
      y1b[p] <- 70
    }
    p=p+1
  }
}

y10g <- y1g - y0g    # calculate the difference of females
y10b <- y1b - y0b    # calculate the difference of males
y0g_mean[i] <- mean(y0g)
y0b_mean[i] <- mean(y0b)
y1g_mean[i] <- mean(y1g)
y1b_mean[i] <- mean(y1b)
y10g_mean[i] <- mean(y10g)
y10b_mean[i] <- mean(y10b)

for (j in 1:NG){
  y0[j] <- y0g[j]
  y1[j] <- y1g[j]
}
for (k in 1:NB){
  y0[k+NG] <- y0b[k]
  y1[k+NG] <- y1b[k]
}
y10 <- y1-y0

##### Analysis part
## propensity score
## get data

```

```

or_data <- data.frame(y0, girl, y1, y10) # the original paradox
or_up <- or_data [ which ( or_data$y0 >= 145),]
## set sub set data <145 pounds
or_low <- or_data [ which ( or_data$y0 < 145),]

library(lme4)
library(Matrix)
library(nonrandom)
# get propensity score
or_up_ps <- pscore(data=or_up, formula = girl~y0, name.pscore = "or.up.ps")
or_low_ps<- pscore(data=or_low, formula = girl~y0, name.pscore = "or.low.ps")

## ps matching
## since in up_weight data, females are less than males, 2 males are matched to one
female
## That means the treated group is set to "1"
or_up_match <- ps.match(object = or_up_ps,
                        ratio = 2, caliper = 0.5, # 2 individual should be matched to
                        givenTmatchingC = TRUE, # male match to female
                        setseed(123456))
or_low_match <- ps.match(object = or_low_ps,
                        ratio = 2, caliper = 0.5, # 2 individual should be matched to
                        givenTmatchingC = FALSE, # female match to male
                        setseed(123456))

## compare matched sample and original data

## balance check using statistical tests
or_up_balance <- ps.balance(object = or_up_match,
                           sel = c ("y0"), #put all the covariates that you want to check
                           method = "stand.diff", alpha = 20)
## balance check Standardized Deviation
or_low_balance <- ps.balance(object = or_low_match,
                             sel = c ("y0"), #put all the covariates that you want to check
                             method = "stand.diff", alpha = 20)

#### Analysis
## Using original data
# Residual gain score
or_reg <- lm(y1 ~ y0 + girl, data= or_data)
or_reg_b[i] <- or_reg$coefficients[3]
or_reg_t[i] <- coef(summary(or_reg)) [3,3]
or_reg_tpv[i] <- coef (summary(or_reg)) [3,4]

```

```

or_up_reg <- lm(y1 ~ y0 + girl, data= or_up)
or_up_reg_b[i] <- or_up_reg$coefficients[3]
or_up_reg_t[i] <- coef(summary(or_up_reg)) [3,3]
or_up_reg_tpv[i] <- coef (summary(or_up_reg)) [3,4]

or_low_reg <- lm(y1 ~ y0 + girl, data= or_low)
or_low_reg_b[i] <- or_low_reg$coefficients[3]
or_low_reg_t[i] <- coef(summary(or_low_reg)) [3,3]
or_low_reg_tpv[i] <- coef (summary(or_low_reg)) [3,4]

# simple gain score
or_up_tt <- t.test(y10 ~ girl, data= or_up)
or_up_diff[i] <- or_up_tt$estimate[2] - or_up_tt$estimate[1]
or_up_t[i] <- or_up_tt$statistic
or_up_tpv[i] <- or_up_tt$p.value

or_low_tt <- t.test(y10 ~ girl, data= or_low)
or_low_diff[i] <- or_low_tt$estimate[2] - or_low_tt$estimate[1]
or_low_t[i] <- or_low_tt$statistic
or_low_tpv[i] <- or_low_tt$p.value

or_tt <- t.test (y10~ girl, data = or_data)
or_diff[i] <- or_tt$estimate[2] - or_tt$estimate[1]
or_t[i] <- or_tt$statistic
or_tpv[i] <- or_tt$p.value

### Using matched data
or.up.matched.data <- or_up_match$data.matched
or.low.matched.data <- or_low_match$data.matched
# Residual gain score
or_up_p.reg <- lm (y1 ~ y0+girl, data = or.up.matched.data)
or_up_p.reg_b[i] <- or_up_p.reg$coefficients[3]
or_up_p.reg_t[i] <- coef(summary(or_up_p.reg)) [3,3]
or_up_p.reg_tpv[i] <- coef (summary(or_up_p.reg)) [3,4]

or_low_p.reg <- lm (y1 ~ y0+girl, data = or.low.matched.data)
or_low_p.reg_b[i] <- or_low_p.reg$coefficients[3]
or_low_p.reg_t[i] <- coef(summary(or_low_p.reg)) [3,3]
or_low_p.reg_tpv[i] <- coef (summary(or_low_p.reg)) [3,4]

# simple gain score

or_low_p.tt <- t.test (y10~ girl, data = or.low.matched.data)
or_low_p.diff[i] <- or_low_p.tt$estimate[2] - or_low_p.tt$estimate[1]
or_low_p.t[i] <- or_low_p.tt$statistic

```

```

or_low_p.tpv[i] <- or_low_p.tt$p.value

or_up_p.tt <- t.test(y10~ girl, data = or.up.matched.data)
or_up_p.diff[i] <- or_up_p.tt$estimate[2] - or_up_p.tt$estimate[1]
or_up_p.t[i] <- or_up_p.tt$statistic
or_up_p.tpv[i] <- or_up_p.tt$p.value

### compared pretest difference
or_upy0_p.tt <- t.test(y0~ girl, data = or.up.matched.data) #matched sample
or_upy0_p.diff[i] <- or_upy0_p.tt$estimate[2] - or_upy0_p.tt$estimate[1]
or_upy0_p.t[i] <- or_upy0_p.tt$statistic
or_upy0_p.tpv[i] <- or_upy0_p.tt$p.value

or_upy0_tt <- t.test(y0~ girl, data = or.up) # original sample up
or_upy0_diff[i] <- or_upy0_tt$estimate[2] - or_upy0_tt$estimate[1]
or_upy0_t[i] <- or_upy0_tt$statistic
or_upy0_tpv[i] <- or_upy0_tt$p.value

or_lowy0_p.tt <- t.test(y0~ girl, data = or.low.matched.data)
or_lowy0_p.diff[i] <- or_lowy0_p.tt$estimate[2] - or_lowy0_p.tt$estimate[1]
or_lowy0_p.t[i] <- or_lowy0_p.tt$statistic
or_lowy0_p.tpv[i] <- or_lowy0_p.tt$p.value

or_lowy0_tt <- t.test(y0~ girl, data = or.low)
or_lowy0_diff[i] <- or_lowy0_tt$estimate[2] - or_lowy0_tt$estimate[1]
or_lowy0_t[i] <- or_lowy0_tt$statistic
or_lowy0_tpv[i] <- or_lowy0_tt$p.value

or_y0_tt <- t.test(y0~ girl, data = or.data)
or_y0_diff[i] <- or_y0_tt$estimate[2] - or_y0_tt$estimate[1]
or_y0_t[i] <- or_y0_tt$statistic
or_y0_tpv[i] <- or_y0_tt$p.value

}

# give output and store
result_data <- data.frame(y0g_mean, y0b_mean, y1g_mean, y1b_mean, y10g_mean,
  y10b_mean, or_diff, or_t, or_tpv, or_reg_b,
  or_reg_t, or_reg_tpv, or_up_diff, or_up_t, or_up_tpv,
  or_up_reg_b, or_up_reg_t, or_up_reg_tpv, or_low_diff, or_low_t,
  or_low_tpv, or_low_reg_b, or_low_reg_t, or_low_reg_tpv,
or_up_p.diff,
  or_up_p.t, or_up_p.tpv, or_up_p.reg_b, or_up_p.reg_t,
or_up_p.reg_tpv,

```



```

        or_low_p.diff, or_low_p.t, or_low_p.tpv, or_low_p.reg_b,
or_low_p.reg_t,
        or_low_p.reg_tpv, or_y0_diff, or_y0_t, or_y0_tpv, or_upy0_p.diff,
        or_upy0_p.t, or_upy0_p.tpv, or_upy0_diff, or_upy0_t, or_upy0_tpv,
        or_lowy0_p.diff, or_lowy0_p.t, or_lowy0_p.tpv, or_lowy0_diff,
or_lowy0_t,
        or_lowy0_tpv)
result_mean <- c(colMeans(result_data, dims = 1))
parameter <- matrix (c(d,back,NN,NT,NG,NB,M0g,M0b,M1g,M1b,SDg,SDb,corr),
        nrow = 1,ncol = 13)
outp<- c (parameter,result_mean)
output[d,]<- outp
assign(paste("data",d),result_data)

# set the data file location in D: and create a data file called "paradox"
if (dir.exists("D:/paradox")) {
  setwd(dir = " D:/paradox ")
} else {
  dir.create("D:/paradox ")
  setwd(dir = " D:/paradox ")
}

# save the simulated data file and result file in current location
filename <- paste("data",d,".txt",sep = "")
write.table(result_data, filename,row.names=FALSE,sep="\t",quote=FALSE)
write.csv(output,"results.csv",row.names=FALSE)
d=d+1
}

```

APPENDIX C

Mplus Syntax for Latent Linear Growth Using Negative Binomial Hurdle Model to Analyze Trajectory Subgroups

Data: file is depr_class1.csv;

variable:

names are idnum diffDepr4_3 diffDepr5_4 MedDepW3 MedDepW4 MedDepW5
TxDepW3 TxDepW4 TxDepW5 depress2 depress3 depress4 depress5 Nzero4_5;

useobservations = (TxDepW3 not EQ 1 or MedDepW3 not EQ 1);
idvariable is idnum;

Missing are all (-9999);

usevariables are

depress2 depress3 depress4;

auxiliary are Nzero4_5;

count are depress2 depress3 depress4 (nbh);

classes= c(3);

Analysis:

type = mixture;

starts = 1000 40;

PROCESSORS =40;

stiterations = 10;

ALGORITHM=INTEGRATION;

estimator = MLR;

Model:

%overall%

i s| depress2@0 depress3@1 depress4@2;

i0 s0 | depress2#1@0 depress3#1@1 depress4#1@2;

i-s@0;

i0-s0@0;

%c#1%

[i@0];

[s@0];

[i0@0] ;

[s0@0];

!depress2@0.01;

depress3@0.01;

!depress4@0.01;

i s| depress2@0 depress3@0 depress4@2;

i0 s0 | depress2#1@0 depress3#1@0 depress4#1@2;

i with i0 (15);

%c#2%

[i] (21);

[s] (22);

[i0] (23);

[s0] (24);

!depress2@0.01;

depress3@0.01;

depress4@0.01;

i with i0 (25);

%c#3%

[i] (31);

[s] (32);

[i0] (33);

[s0] (34);

!depress2@0.01;

i WITH i0 (35);

output: stdyx tech1 tech11 ! tech14 ! tech4;

savedata: file is hurdle3_g3.csv;

save is cprob;

APPENDIX D

R Syntax for Multiple Imputation on Missing Data in FFCW Data Set

```
rm(list=ls())
setwd("D:/paradox")
## get the main data: IV, DV, and COV
library(haven)
## Use original data
pscore <- readRDS("R/pscore_variab_raw.rds")
pscore[] <- lapply(pscore, unclass)

##### mutiple imputation using the package mice
pscoeMiss <- function(x){sum(is.na(x))/length(x)*100}
apply(pscore, 2, pscoeMiss)
library(mice)
md.pattern(pscore)
library(VIM)
pscore_plot <- aggr(pscore, col=c('navyblue','red'), numbers=TRUE,
                    sortVars=TRUE, labels=names(data), cex.axis=.7,
                    gap=3, ylab=c("Histogram of missing data","Pattern"))
pscore.imputing <- mice(pscore, m = 10, maxit = 50, method = "pmm", seed = 500)
xyplot(pscore.imputing, depress5 ~ depress4, pch=18,cex=1)
densityplot(pscore.imputing)

pscore.imputed <- complete(pscore.imputing)

pscore.imputing <- readRDS("R/pscore_variab_imputing.rds")
# iv.dv <- subset(pscore, select = "idnum", "depress5", "TxDepW4", "MedDepW4")
saveRDS(pscore.imputed, "R/pscore_variab_imputed.rds")
saveRDS(pscore.imputing, "R/pscore_variab_imputing.rds")

pscore.imputing <- readRDS("R/pscore_variab_imputing.rds")
```

VITA

Hua Lin

Candidate for the Degree of

Doctor of Philosophy

Dissertation: REVEALING AND RESOLVING CONTRADICTORY WAYS TO
REDUCE SELECTION BIAS TO ENHANCE THE VALIDITY OF
CAUSAL INFERENCES FROM NON_RANDOMIZED
LONGITUDINAL DATA

Major Field: Human Development and Family Science

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Human
Development and Family Science at Oklahoma State University, Stillwater,
Oklahoma in July, 2018.

Completed the requirements for the Master of Science in Human Development
and Family Science at Oklahoma State University, Stillwater, Oklahoma in July,
2015.

Completed the requirements for the Bachelor of Science in Applied Physics at
Jinan University, Guangzhou P. R. China in 2001.

Experience:

Instructor, Graduate Teaching Associate, Graduate Research Associate, Human
Development and Family Science, Oklahoma State University,
Stillwater, Oklahoma.

Research/Teaching Customs Valuation and Information, Guangzhou
Merchandise Valuation and Information Office of General
Administration of Customs, Guangzhou, P. R. China.

Professional Memberships: